

**Draft Report of the National Toxicology Program Board of
Scientific Counselors Working Group on Evaluating NTP's
Approach for Reaching Conclusions for Literature-Based
Evidence Assessments**

December 11, 2012

Submitted by:
Lynn Goldman, MD (chair)
Reeder Sams II, PhD (vice-chair)

Preface

The National Toxicology Program (NTP) Board of Scientific Counselors (BSC) is a federally chartered external advisory group that provides input on the scientific merit of NTP's programs and activities. The NTP BSC Working Group (WG) on Reaching Evidence Assessment Conclusions was formed in July 2012. The purpose of the WG was to evaluate the *Draft NTP Approach for Reaching Conclusions for Literature-Based Evidence Assessments* (the *Approach*), which is a proposed methodology to use the information gathered by a transparent systematic review process to reach hazard identification conclusions.

Overall, the WG commended the NTP for taking proactive steps to increase the transparency of hazard assessments. The WG enthusiastically supported the development of the *Approach*. NTP's development of the methodology reviewed by this WG is consistently moving forward the state-of-the-science for hazard assessment and is responsive to recent recommendations from authoritative scientific organizations (e.g., National Academies of Science, NAS). The Working Group encourages NTP to advance and evolve methodologies for hazard assessment.

The WG was composed of 10 scientists representing academia, industry, and government. Dr. Lynn Goldman, Dean and Professor, School of Public Health & Health Services, George Washington University, chaired the WG. Dr. Reeder Sams, Senior Science Advisor, National Center for Environmental Assessment/RTP Division, U.S. Environmental Protection Agency (EPA) served as vice-chair. The WG roster is attached (Appendix A). National Institute of Environmental Health Sciences (NIEHS)/NTP staff attending the meeting were Drs. John Bucher, Mary Wolfe, Kristina Thayer, Andrew Rooney, Abee Boyles, and Lori White (NTP BSC Designated Federal Official). The WG met August 28 and 29, 2012, at the Raleigh Marriott Crabtree Valley, 4500 Marriott Drive, Raleigh, NC.

As background for the meeting, the WG members were provided with the *Approach* (Appendix B) and the format for the meeting (Appendix C). Dr. Thayer, Director of NTP's Office of Hazard Assessment and Translation (OHAT), opened the meeting by providing an overview of the development of the draft NTP approach and how it fits within the OHAT evaluation process. Dr. Rooney, Deputy Director of OHAT, then provided background on the approach and the charge:

to obtain feedback on the NTP's proposed approach for reaching conclusions for literature-based evidence assessments.

Drs. Rooney and Boyles made presentations on the specific areas for working group consultation as per the meeting format. The WG discussed each step of the approach and responded to the questions posed in the meeting format.

Description of the NTP BSC WG Report Format

The following report summarizes discussions during the face-to-face meeting of the NTP BSC WG on August 28 and 29, 2012. The WG was charged with reviewing the *Approach*. The draft NTP document describing their proposed approach was divided into seven steps. The discussion held by the WG focused on each of the seven steps and the following report is organized to coincide with these steps. For each step the WG has prepared the following components (1) a brief overview capturing NTP's proposed approach and the focus of the discussion, (2) recommendations for which the WG achieved consensus, and (3) comments that were made during the WG discussions that may provide useful information to NTP as they move forward to develop their methods for hazard assessment. Please note, the comments that are recorded for each step the section entitled *Specific WG comments for consideration by NTP* do not represent a recommendation or a consensus opinion of the WG. Some comments may represent a minority or divergent opinion of some WG members. They are presented to provide the BSC a complete picture of the WG discussions that led to our recommendations.

Step 1: Prepare Topic

WG Discussion Overview:

Step 1 of the *Approach* provides an approach to prepare a focused topic and draft protocol for developing a hazard assessment. As described by NTP, Step 1 documents the specific approach to be used in the evaluation including key aspects of searching for and selecting studies, grouping related outcomes, forms for data extraction, and how risk of bias will be assessed. The NTP BSC WG recommendations regarding this step include the following:

WG recommendations:

- 1) For each substance being evaluated, the NTP should establish a draft protocol including risk of bias (RoB) questions *a priori*.
- 2) Development of the draft protocol for evaluating a substance should follow an iterative process and be refined based upon subsequent steps (up to Step 4) and become immutable (except in rare documented circumstances) prior to reaching evidence-based conclusions on hazard assessment.

- 3) Consideration of relevant human exposure levels including specific and susceptible populations should inform the scope of the topic.
- 4) The NTP should consider a broad spectrum of scientific information to clearly reach decisions regarding study design or preparing the topic for a hazard assessment.

Specific WG comments for consideration by NTP:

- a) Overall, Step 1, *Prepare the Topic*, needs to be clear, transparent, and adhere to sound scientific principles with respect to the study design. Populations, exposures, and outcomes of interest should be specified within this step. The test system should include PICO/PECO principles (population, intervention or exposure, control or comparator, and outcome) and risk of bias assessment criteria.
- b) Hypothetically, for or any given topic investigated by NTP, there will exist a wide range of available data within the spectrum of relevant to irrelevant. NTP should conserve resources and focus on data that are relevant to health outcomes and useful for elucidating key events that lead to health effects in humans.
- c) Early in the process (i.e., Step 1) it is important to consider the impact of relevant human exposure levels for defining the topic and ultimately for the utility of hazard assessment conclusions. Additionally, the NTP should be aware of the exposure ranges that may be relevant to various populations (i.e., the general population, children and women of childbearing age, consumers, workers, and people in more highly exposed communities). Although the draft NTP methodologies reviewed by the WG were not intended for exposure studies, NTP may wish to develop a process for evaluating such studies and how exposure information would inform hazard assessment conclusions.
- d) The NTP should consider the entire spectrum of experimental evidence (e.g., structure-activity-relationships [SAR], mechanisms, mode of action, kinetics, information from related toxicants, etc.) in addition to human epidemiological and toxicological animal studies to clearly reach decisions with respect to the study design or preparing the topic for an assessment.
- e) For systematic review, developing the hypothesis(es) and defining the questions are critically important and must be decided *a priori*. Failure to adhere to these

principles can lead to spurious results and conclusions. Additionally, there exists a concern with pre-specification of preparing the topic / hazard questions. The use of an adaptive design would address the concern of not identifying potential outcomes. Utilizing this type of an approach would allow NTP to determine if sufficient evidence exists to proceed with a hazard evaluation. Alternatively, if sufficient evidence is not identified, the NTP could utilize an adaptive design to broaden the scope of the topic to determine if additional data would inform the assessment. An adaptive design would conserve resources by potentially eliminating the need to identify data that are not apparently relevant to the scope of the assessment and prevent bias identifying health outcomes at the onset of the systematic review. The NTP should not lock in prematurely on a protocol for conducting an assessment in Step 1; an iterative determination should be made at step 4 (*Assess the Quality of Individual Studies*), where decisions are made about what additional data are brought in based on the data extraction at that point. Once those other studies are brought in, they need to be assessed with a high degree of rigor, and as the topic is redefined, the process can become immutable. However, it is important to utilize focused questions to obtain relevant data and conserve resources.

- f) The protocol provided by the NTP to the WG defines the questions and analyses and it is important not to decrease credibility by performing *post hoc* sub-analysis, resulting in the removal of a health outcome or by introducing new outcomes in the later stages of the process. Multiple protocols can be developed to address related hazard questions and applied to a comprehensive data set at a later date.
- g) A variety of public health decisions need to be made on data-light chemicals (i.e., small data sets). The NTP review framework should be suitable for these chemicals although more focused questions may be required for which there may be implications in later steps in the process. With data-light chemicals, there may exist a greater risk of study sponsor publication bias. On the other hand, data-light chemicals increase the need and opportunity to consider information from putatively related chemicals or chemicals with putatively similar modes of action.
- h) The NTP should utilize agency partners and other experts to frame the questions for Step 1.
- i) Where SAR, physiologically based pharmacokinetic (PBPK) models, and exposure assessment data are salient to the study design or preparing the topic

for an assessment, it is important to systematically assess the quality of such studies with the same high standards applied to hazard assessment studies. The NTP may wish to develop a tool to assess these types of studies.

- j) Relevant to the scope of an assessment, NTP should develop a process for determining relevant health outcomes. Health outcomes should include a broad array of scientific information (e.g., biomarkers of effect, upstream biological endpoints, etc.), not just traditional toxicological endpoints or apical outcomes from human studies. A transparent process should be employed to identify health outcomes. This stage of assessment development should include hypotheses focused on the grouping of related health outcomes.
- k) There is a need to develop guidance for use of data collection sheets to get robustness and there is a need for transparency in the data collection. There is a need for multiple data sheets that can be modified for individual substances for the specific questions that are being asked. The data forms will likely evolve as assessments are completed.
- l) The NTP process should develop an avenue to consider relevant studies that are published during the development of an assessment as well as a potential shift in the state of the science regarding the interpretation of a body of literature (i.e., assessments can take years, so as new studies are published, there needs to be a way to incorporate them using updated searches and new interpretation of important papers).
- m) Identifying the end-user (i.e., client or stakeholder) needs (e.g., regulatory action), will assist in determining the scope of the assessment. Similarly, information regarding the assessment development and conclusions should be transparently communicated to stakeholders and the public.

Step 2: Search for and Select Studies for Inclusion

Discussion Overview:

Step 2 of the *Approach* focuses on identifying and selecting studies to undergo data extraction (Step 3). The draft approach specifies the need to screen studies based upon inclusion / exclusion criteria and briefly describes the logistics pertaining to the implementation of Step 2.

WG recommendations:

- 1) NTP should conduct a thorough literature search for all studies relevant to human health for a given topic or hazard assessment. The literature search strategy should be transparently described in NTP's hazard assessment documents.
- 2) Studies utilized for subsequent steps of systematic review should be independently peer reviewed. If they have not yet been peer reviewed, the NTP should arrange for peer review, utilizing existing NTP processes to conduct independent peer reviews.
- 3) Specific data types (e.g., pharmacokinetic (PK), *in vitro*) data should be identified with respect to informing the hazard assessment.

Specific WG comments for consideration by NTP:

- a) A thorough literature search should be conducted for all studies, although it may be difficult to avoid publication bias. Additionally, NTP should provide opportunities for agency partners and stakeholders to submit unpublished studies that would inform hazard assessment.
- b) The use of gray literature in systematic reviews should be transparent and consistent with NTP policy (i.e., conduct independent peer-reviews for non peer-reviewed literature). Narrative reviews are more prone to author-bias, not likely to add significant information for the development of an assessment, nor are they likely to contain original research which would meet the criteria for inclusion in systematic review. On a case-by-case basis there may be the need to request additional information from research organizations (including stakeholders), which may require additional peer review. However, if the additional information involves provision of simple statistics that were not reported (e.g., standard

deviation, sample size) additional peer review may not be required. Processes for evaluating additional data should be documented in a transparent manner.

- c) When available, PK data should be considered during the development of an assessment. PK studies should be used where the NTP needs to perform dose translation across species, across dose groups, and in quantification of inter-individual variability. PK data often are useful to assess quality of individual studies, consistency across outcomes, and population groups (e.g., lifestage, chronic disease, and other susceptible populations). PK data may help to inform which exposure measures are most relevant. There needs to be knowledge of acceptable exposure measurements, confidence in the exposure assessment, and use of judgment for effective use of PK data. Thus, how and where PK data are utilized and inform the hazard assessment should be addressed in the protocol.
- d) The WG encourages the NTP to advance the methodology for utilizing *in vitro* studies in hazard assessment. There is a need to look at the utility of *in vitro* data in hazard identification. In assessing the utility of *in vitro* data, human relevance is the first consideration, however *in vitro* data may inform about surrogate markers, may provide information that helps extrapolate hazard data across species, and may inform about critical issues such as routes of exposure, life stage-specific events, and biologic plausibility. There are a lot of poor quality or irrelevant *in vitro* data and it is a challenge to work through them hierarchically. *In vitro* studies can be problematic in terms of the *in vitro* signal as it relates to outcome measures and in terms of dose and rate issues that are hard to interpret. Mutagenicity data are also sometimes hard to incorporate because of quality assessment issues and differential responses across various platforms. Generic things to look for in *in vitro* studies are enzyme inhibition, receptor binding, and other endpoints that affect the signal. It was suggested to rely on Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM)-validated assays, which are part of a regulatory framework. However, it was agreed that many non-ICCVAM-validated assays may provide valuable information and NTP should consider pre-established criteria for inclusion and review of these studies. Criteria should focus on quality and relevance of *in vitro* studies to human hazards.
- e) Importantly is how to frame the question and where to put in inclusion/exclusion criteria.

- f) The NTP could use an iterative approach where human data are evaluated first, and if they provide sufficient evidence of human hazard, it is not necessary to use animal data in reaching a primary conclusion of hazard. This approach provides efficiency and uses judgment at the point of preparing the topic and the prior knowledge that a hazard is present. However, animal and other data may still be relevant for assessments of likely mode of action, and associated conclusions about likely forms for dose response relationships, and identification of likely sensitive subgroups of the human population.
- g) There are two worlds of chemicals, those that have been studied for decades (e.g., arsenic, DDT, lead) and those for which there are very little data. NTP needs to be rational in using *in vitro* data for the well-characterized chemicals.
- h) Model systems should be evaluated and decisions should be made in the study designs about where they fit in. Not all animal models are relevant for assessing human health effects, e.g., it was suggested that *C. elegans* data might be relevant only to mechanistic information/mode of action.

Step 3: Extract Data from Studies

Discussion Overview:

Step 3 of the *Approach* specifies the approach to extract study data. Appendices to NTP's draft approach provided example data extraction forms for humans, animals, and meta-analysis studies. General logistics for how data extraction would be performed was also presented in the draft approach. The WG made two specific recommendations for Step 3.

WG recommendations:

- 1) The NTP should utilize separate data extraction frameworks for animal and human datasets and not be overly concerned with a homogeneous approach for different types of datasets.
- 2) Data entry should be quality controlled, for example, via conduct by two independent data extractors.
- 3) Initially, the NTP should incorporate more weight to non-apical studies.

Specific WG comments for consideration by NTP:

- a) For data extraction it appears there is an attempt to force fit experimental animal studies into an epidemiology framework; reconsider using common terminology for those two streams of evidence. Some terms may have different definitions when used in experimental animal and human epidemiological studies; for example the word *attrition*. Certainly the WG agrees that experimental animal studies should more consistently report dropouts because this information is important in understanding studies' findings. However, it is important to recognize historically, that animal studies have not consistently reported dropouts. How data on attrition/dropouts in experimental animal studies will be handled should be defined in the individual protocol.
- b) The WG had support for using two independent reviewers for an evaluation and having a system for conflict resolution.
- c) The *Approach* should clarify the difference between studies and papers, e.g., one study can be reported in several papers and a single paper can report about several studies. Also, one cohort (or experimental animal study) can be reported in multiple studies and such studies should not be considered independent investigations.

- d) As noted previously, the WG would like the NTP to consider non-apical studies (e.g., *in vitro* studies) from the outset of the review process. The study design needs to consider which endpoints can be evaluated; especially in human studies some endpoints are not captured because they cannot be done ethically. The design needs to factor in considerations of sensitivity and the ability to capture more subtle endpoints that may be detectable with statistical confidence at lower dose levels and may help understand the entire continuum of adverse effects.
- e) As noted above, the WG considers it important to consider broadly all evidence streams. Human evidence evaluations should be based on the scientific rigor of the studies, irrespective of biologic plausibility and priors; however, other relevant evidence like kinetics and validation studies of biomarkers need to be considered and that evidence in turn needs to undergo a quality review.

Step 4: Assess the Quality or Risk of Bias of Individual Studies

Discussion Overview:

NTP outlined an approach for assessing the quality of individual studies, carefully defining it to apply to internal study validity or risk of bias (RoB). Within this step, the methods developed by NTP were based on methods previously developed by the Agency for Healthcare Research and Quality (AHRQ; Viswanathan *et al.*, 2012). The approach included 18 specific RoB questions, of which a specific subset of the questions for a given study design are used to ascertain RoB for an individual study (e.g., a different subset of the 18 questions is asked for cohort studies, than for case-control studies). The NTP proposed to designate four RoB questions as having greater importance for ascertaining the confidence that exposure to environmental substances are associated with health effects and designated these questions as *major* RoB questions as a means to exclude studies that are judged to have “Definitely” high RoB for two of the major questions.

WG recommendations:

- 1) The approach outlined by the NTP to evaluate study quality or RoB of individual studies is reasonable and supported by the WG.
- 2) The 18 RoB questions could be worded more plainly to address study quality with language that aids in the interpretation of the question for animal studies. The WG suggested rewording the questions.
- 3) The WG suggested dropping the designation of a subset of questions as *major* risk of bias questions. The WG recognized NTP was attempting to use the *major* questions as a means of excluding lower quality studies as the basis for conclusions. The WG was split on the question of excluding studies. However, the WG did not support a pre-defined subset of RoB questions as being more definitive compared to other questions or to use these pre-defined subset of *major* RoB to exclude studies for every systematic review that might be undertaken by the NTP.

Specific WG comments for consideration by NTP:

- a) In fact, it was noted that the term *RoB* is not understood by people who are not epidemiologists; therefore, the process and terminology will require transparent communication. In general, the wording for many of the questions was not

applicable or relevant to animal studies. Some WG members were concerned that this could create a bias against animal studies because they felt many studies would unjustifiably be downgraded based on RoB; other WG members were concerned that this would result in RoB in animal studies going undetected or characterized inappropriately by reviewers. With regard to the first concern, some of the information requested in the RoB questions is not routinely reported in animal studies, which may lead to inappropriate downgrading due to high bias. Also, toxicology studies often have a number of missing endpoints. *Not reported* is different from *probably high risk of bias*. It is not always possible to contact the authors to get additional information. On the other hand, other WG members thought that animal studies have the same obligation as human studies to report on factors related to RoB and that these factors could contribute to RoB. There was WG consensus, however, that the NTP should limit exclusion of animal data.

- b) As noted above, the WG agreed that the RoB questions should be clear about the factors that are likely to cause RoB for experimental animal studies. Too often in the view of the WG, reviews only assess whether or not a study was conducted under GLP regulatory standards and do not provide a detailed analysis of the factors that can lead to bias in animal studies. Many useful studies are not conducted in GLP-certified labs; this does not mean that they don't use procedures enshrined in GLP regulatory standards. While GLP does not indicate a higher rated study, it does mean there is greater documentation and adherence to agreed-upon guidelines. The NTP should explicitly identify factors that can lead to bias in animal studies, e.g., control for litter effects, dosing, and methods for exposure assessment, so that these factors are consistently evaluated for experimental studies.
- c) Likewise the WG agreed that the RoB questions needed to be reworded so that they are relevant to cross-sectional human studies and non-experimental animal studies, as well as other studies that may be relevant including *in vitro* studies and PK studies. Some of the WG suggested that a separate set of RoB questions might be required for differing study designs.
- d) Regarding detection in animal and in human studies, the question posed was whether the adverse effect occurred after treatment. The WG suggested that not just the ordering of events (whether the exposure preceded the outcome), but also timing should be considered. For example, some effects require a latency period (e.g. carcinogenesis) and others only occur with certain exposure windows (e.g. congenital malformations).

- e) The person doing the coding should have guidelines for using the questions. There should be backup documentation.
- f) Kinetic and other types of data should be included and reviewed at the same time to guard against bias; it is more efficient to do it that way.
- g) In publications with more than one type of data (or several studies embedded within them), there needs to be an evaluation of RoB for each type of data.
- h) The WG did not have a unanimous view regarding the *ideal* number of categories in the RoB ranking scale. Generally it was felt that a 3-point scale does make it too easy for reviewers to gravitate toward the middle (*moderate*) level. It was noted that the Cochrane Collaboration's use of an intermediate category has resulted in high use of the middle category; it has used a 3-point scale but is considering moving to a 4-point scale for this very reason.
- i) The WG did not have a unanimous view regarding quantification of study RoB. Generally, the WG thought that such quantification has the potential for giving an appearance of precision where the ability does not exist to quantify RoB. However, one member thought that RoB should be quantified. All agreed that there is a need for more research to understand the influence of RoB quantitatively, and that information relevant to RoB should be collected for subsequent evaluations. There is a need for a fair, consistent, and transparent process for making judgments about RoB for each study.
- j) It was suggested that another important RoB question is: Was the animal model appropriate? However, a concern was raised as to whether criteria for this RoB question could be defined *a priori*.
- k) Blinding has been shown to affect outcomes, but it is poorly reported. Generally, the WG thought that animal studies with blinding are higher quality studies but there was not consensus on this point. One WG member pointed out that in the case of experimental animal studies, treatment groups may not always be blinded and sometimes the identity of the treatment group is obvious even if one makes an attempt to blind the observers.
- l) The WG agreed that randomization in experimental animal studies is an important RoB factor. However, randomization is often not reported in toxicology studies, because it was not done, because the researchers did not think it was important, and/or because toxicology journals do not require it to be reported.

On this basis, there was not a unanimous WG view regarding randomization as a RoB factor. While several such factors (e.g., randomization, attrition) have been found to increase RoB in clinical studies, members of the committee were not aware of any RoB analyses from toxicology studies.

- m) Contrary to what was suggested in the draft *Approach*, studies incorporating *emerging* endpoints and biomarkers may have a higher RoB. The new studies often have not been replicated and there is more uncertainty regarding the endpoints. Some WG members expressed the opinion that there may be more risk of investigator bias in these situations (as has been documented in the case of pioneering medical interventions and pharmaceuticals). Novel assays need to be reviewed on a case-by-case basis and reviewers need to be trained on how to evaluate new studies. Studies with novel endpoints should be incorporated early in the protocol and the process by which RoB will be evaluated should be determined *a priori*.
- n) The relative importance of the RoB questions can be determined only after a number of reviews have been done. There is no need to designate *major* RoB questions. The draft *Approach* posits that if there are dark reds (*high* RoB) for two of the four *major* questions, the study should be excluded. The WG did not have a consensus that this is the case and in fact were fairly evenly divided over the question of whether any studies should be excluded at this phase. The WG agreed that some questions are more important than others, but concluded that study results can be ordered or ranked (or weighted differently) based on confidence of studies. Importantly, bias can sometimes be dealt with by utilizing sensitivity analyses.
- o) Some WG members expressed the opinion that studies can be evaluated or grouped by different RoB scores to assess the influence of various RoB factors. Generally, the WG concurred that it is important to exclude studies with obvious fatal flaws and there needs to be a transparent process and clear criteria for the NTP to exclude studies with fatal flaws. Examples of such criteria include: instability of test compound, or inadequate or no controls (or comparison group), or unreliable measure of exposure or outcome. The process for excluding studies should have been determined at the protocol development stage. However, a few WG members thought it important to get the question right so “good” studies are not excluded. Importantly, one WG member noted that the GRADE process does not exclude study findings, but rather ranks them and allows reviewers to make their own judgments.

- p) The question, “*Does the study design adjust/control for important confounding and modifying variables?*” needs to be reworded; use of the word “important” is unclear. As written, the question is not directly relevant to animal studies. There needs to be a different question to address design of experimental animal studies or this question needs to be rewritten to clearly include issues relative to both epidemiology and animal studies.
- q) Statistical power needs to be incorporated into the list of RoB questions. Underpowered studies are a very serious source of bias (towards the null) and a huge problem in epidemiology studies. This is a critical issue. Lack of power is related to inadequate numbers of study subjects, inadequate exposure differences, or both. Along with assessing numbers of subjects, for epidemiology studies insufficient spread in exposure, which is similar to insufficient dose ranges in animal studies, should be addressed in the RoB questions.
- r) Inaccuracy of exposure measures is a crucial issue. It is important to know what the uncertainties are and how to counteract them in later quantitative analysis.
- s) The NTP needs to take care in how it discusses potential confounding as a RoB issue. In statistical analysis, throwing a lot of variables into a multivariate model is not the same as controlling for confounding. For a reviewer to make a judgment that there is lack of control for confounding, the study design should have established *a priori* which factors are potential confounders and need to be addressed.
- t) It is important to not have a rule to always exclude studies on account of bias. In some cases, biases can be handled later in the analysis, in others not. RoB elements need to be empirically based. For example, a common bias in occupational epidemiology studies is the so-called “healthy worker effect” in which a working population which is intrinsically healthier is inappropriately compared with general population controls, who have a much greater baseline disease burden. Protocols should detail how to identify and deal with the healthy worker effect bias, which is so well understood that it was suggested that it might be possible to use empirical analyses to offset it. The protocol should include how to incorporate related positive controls to show an effect and build it into a quality assessment. It is difficult to get a positive affirmation of a negative effect; authors often need to be contacted
- u) Some WG member expressed the opinion that epidemiological studies observing no exposure-related effects due to a pollutant are currently not adequately

considered in hazard assessment. However, negative studies should only be considered with the caveat that adequate study design and power were employed in the conduct of the study, which would be largely addressed by RoB.

Step 5: Rate the Confidence in the Body of Evidence

Discussion Overview:

The NTP presented an approach for developing confidence ratings for the collection of studies or body of evidence. The approach started with definitions for 4 levels of confidence from High confidence to Very Low confidence and initial confidence ratings by study design type. The NTP outlined factors that would lead to downgrading as well as upgrading the confidence rating for the body of evidence. The NTP specifically asked the WG to consider whether consistency across study designs, across populations, and between species should be considered as additional factors to increase confidence in the association between exposure to a substance and a health outcome. The NTP also outlined a procedure for developing confidence ratings across multiple study types and multiple effects or endpoints.

WG recommendations:

- 1) The WG suggested several changes to initial confidence ratings:
 - a. That the term *ecological studies* be removed from consideration as a study type for initial confidence rating (*ecological* refers to exposure classification, not a study type).
 - b. The WG suggested that caution should be used when evaluating the initial confidence for case-reports as they could be used as the basis for important public health decisions, depending on the study question.
 - c. The WG suggested that case-control and nested case-control studies could be given the same initial confidence rating as cohort studies because there are high quality case-control and nested case-control studies that are comparable to cohort studies.
- 2) The WG suggested that some of the reasons for downgrading confidence in the body of evidence should be explained in greater detail. In comparison to RoB, the issues were not thoroughly described.
- 3) The WG supported the NTP's list of factors that could decrease confidence in a body of evidence and the factors that could increase confidence in a body of evidence. Specifically, the WG agreed that consistency across study designs, populations, and species should be part of the NTP's list of factors that could increase confidence in a body of evidence. In addition, the WG suggested

adding consideration of *rare outcomes*, *harm*, and *specificity* as factors that could increase confidence in a body of evidence.

Specific WG comments for consideration by NTP:

- a) Nested case-control studies and case-control studies should not automatically be ranked lower than cohort studies, nor should case-control studies be automatically ranked below nested case control studies. Ideally, the quality of case-control studies can approximate the quality of nested case-control studies and cohort studies. Decisions about whether findings from case-control studies can be upgraded should be decided in the protocol and can include the appropriateness of controls, the timing of exposure measures with respect to the onset of disease, and other design factors rather than judging solely on the basis of the study design. For example, a case-control study utilizing population-based controls and earlier measures of exposure (from archived samples or clinical records) may not be more biased than a cohort study. A case-control study with clinical controls and utilizing recall for exposure assessment would potentially be more biased and would be ranked lower.
- b) The term *ecological studies* is not useful in this context. Rather than classifying studies as *ecological* (or not) it is important to critically assess the use of ecological measures and whether such measures are appropriate in the context of the study or may create bias (the *ecological fallacy*) in the case of measures that are too broad or likely to be confounded by other unmeasured exposures.
- c) The WG saw little value in simply counting studies of various types. It is key to evaluate the toxicity of the substance, not the body of evidence. The focus should be on the identification of higher quality studies that can drive the process of making judgment calls. Such higher quality studies shouldn't be discounted due to larger numbers of poor quality studies.
- d) The draft *Approach* proposes that case-reports initially receive a low ranking that prevents them from getting a high confidence. The WG agreed that the value of case-reports might not be readily apparent to the reviewer. For certain health endpoints, e.g., acute pesticide poisoning, they can provide evidence for a health hazard. Case-reports may not provide the best data for carcinogenicity evaluations; however, even in the case of cancer, rarely have case series provided important information (e.g., DES and vaginal cancer, vinyl chloride and

angiosarcoma of the liver). In such cases, case report data in combination with animal studies have been useful for hazard identification.

- e) Upgrading and downgrading evidence based on strengths and concerns is a good approach. There is no need to worry about balancing the numbers of factors that decrease or increase confidence, but it is important to capture all the significant factors that can upgrade or downgrade a body of evidence, for the sake of fuller transparency in how reviewers are drawing conclusions.
- f) There are statistical issues. Inadequate power needs to be included as a factor that decreases confidence, or it can be included in one of the study design-specific RoB questions. Meta-analysis or pooling of data can be used to combine results from several low-powered studies. Thus, it is not wise to base decisions on statistical significance alone, and the reliance on *p-value* < 0.05 as an indicator of a strong finding has been overemphasized.
- g) Consistency should be added to the factors that increase confidence. Consistency among studies should be evaluated in more than one place to account for health effect-driven processes and fit within biological expectations. When adding consistency (across species or study design) as one of the factors for increasing confidence, it is possible that the confidence in a body of evidence can be elevated from moderate to high. The WG agreed that it is important to explore factors that might explain inconsistency in a body of literature, e.g., various dose ranges, species differences, variability in study quality, different study sponsors, and different study laboratories. If there are inconsistencies, one can do a sensitivity analysis or can try to remove the inconsistencies by arraying by dose, species, and study type. It should be made clear in the schematic that inconsistency is determined for a body of evidence, whereas RoB evaluates internal validity of a single study.
- h) There should be additional factors included that allow for increasing confidence such as mechanistic data, PK data, mode of action, consistency across related biological outcomes, consistency across species, and internal consistency of findings in a study.
- i) Effect size is very outcome dependent. It is important to use a standardized effect size, which should map across different outcome measures. With small samples, the measured variance is only a reflection of the population variance.

- j) Some WG members suggested for the phrase *all plausible confounding*, the word *plausible* should not be used. Consider using *important*, *relevant*, *impactful*, or *reasonable*.
- k) The WG supported the statement: *The confidence conclusions for biologically related outcomes may be assessed for each outcome separately and then reassessed after combining data for all the related outcomes. The overall confidence rating for combined outcomes can differ from that of the individual outcome ratings.*
- l) Regarding the statement: *There are more potential factors by which confidence in the body of evidence can be decreased than increased. Is this approach balanced, or is it imbalanced considering a health protective goal?* The phrase, *considering a health protective goal*, should be removed.
- m) An additional step should be added in which there can be integration across the previous steps to consider consistency across outcome, species, etc. This will provide a way to deal with low ranking conclusions and a rationale for upgrading the conclusions.
- n) Likewise it is not useful to determine a confidence level for all experimental animal studies as a single group. There is tremendous variability in quality among experimental animal studies. One viewpoint expressed among WG members was that animal studies should be ranked high initially because they are experimental and not observational; there is great control due to the animals being identical and they give high quality information. Other WG members did not agree with this position.
- o) A large question is how to make decisions to not use poor quality studies and those that do not contribute to the bottom line conclusions. As noted above, the WG did not have consensus on the process for dropping poor quality studies. Some felt that including the whole body of literature muddies the water, causes decreased confidence in the data, and has implications for evaluating the dose response. All agreed there is the need to avoid the appearance of cherry picking studies and to explain carefully why studies are excluded. Many felt that rather than dropping studies, poorer quality studies can be retained and later determined to be supportive or not supportive. Alternatively a formal meta-analysis that incorporates all studies can be used to evaluate the data.

- p) In any case, determining criteria *a priori* in the protocol is a way to eliminate the incorporation of poor quality research that either should be dropped, used only as supportive (or not) studies, or down-weighted in a meta-analysis. For example, one could *a priori* decide to exclude mortality data for an assessment of an outcome that is best assessed with incidence data because of treatment or reporting biases.
- q) To determine criteria for exclusion or down-weighting evidence *a priori* in the protocol, it is dependent on adequately defining the topic to address appropriate hazard assessment questions. It is important that the NTP more strategically deploys its resources and does more assessments to identify the potential hazards of substances. This argues for more of a focus on the best evidence and less time spent on an unproductive review / analysis of less informative evidence.
- r) Imprecision (wide confidence intervals) may decrease confidence, but one WG member made the helpful suggestion that imprecision often can be addressed quantitatively in the study design using a Bayesian analytic approach to assessing the data.
- s) It is important to consider the consistency of the dose response across studies. However, a dose-response curve is not necessary for each individual study.
- t) In judging the need to control for confounders, consider calculating the extent to which the potential confounding factors for a relationship are likely to modify a relative risk.
- u) In experimental animal studies, the magnitude of effect should not be considered purely from the viewpoint of the associated *p*-value. Instead consider using benchmark dose response to identify relevant biological responses.
- v) There should be a discussion of the evidence that can be used by risk managers.
- w) Regarding indirectness as a factor in decreasing confidence, there was concern that care should be taken before downgrading a study because one may miss biological indicators, either basal or apical. This could possibly be dealt with in the study design by identifying the direct measures *a priori* and then giving less weight to indirect measures. To do so requires knowledge of whether events are likely to be causally related (e.g., “biologic plausibility”). It’s important to be clear and transparent regarding indirectness to allow for scientific discourse.

Step 6: Translating Confidence Rating into Evidence of Health Effects

Discussion Overview:

The NTP detailed an approach for translating confidence in the body of evidence into evidence of a health effect by considering the confidence in the findings with respect to whether or not exposure is associated with a health outcome (i.e., toxicity or no toxicity).

WG recommendation:

- 1) The WG suggested changing the terms used to describe evidence of a health effect or the descriptors. While *sufficient* was an acceptable term, *limited*, and *inadequate* had connotations that make the terms problematic for describing a set of studies that could then move forward as the basis for conclusions.

Specific WG comments for consideration by NTP:

- a) The WG suggested that the NTP consider adding another category, termed *not assessed*, for health effects for which there are no data, e.g., today asthma is not assessed in rodent models. (As noted above, specification of which endpoints are assessed should be addressed initially in the protocol.) Some WG members suggested adding an additional new category intermediate between *sufficient* and *limited*. Also there was concern about the descriptors for these categories, i.e., the words *sufficient* and *limited*. The NTP clarified that the goal was to create a database that other agencies can use with their own terminology.
- b) The WG suggested that NTP use the terms *high*, *moderate*, *low*, *very low* instead. It is very hard to prove *no effect*, so many ratings end up calling the evidence *inadequate*.
- c) The point was made that there can be two ways to categorize the evidence when no effect is found: *inadequate evidence of a health effect* or *inadequate evidence of no health effect*. It can make a big difference in terms of evidence of harm and evidence of no harm; there is an attempt to put them on the same scale. It would be better to state it more neutrally as: *inadequate evidence to draw a conclusion*.

- d) The WG agreed that a conclusion of *evidence of no health effect* requires high confidence in the body of evidence.
- e) The WG also agreed that the outcome with highest evidence of a health effect conclusion moves forward for hazard identification but that all outcomes should move forward.
- f) There needs to be guidance for dealing with very small databases which may have only one really good study showing no effect (e.g., some NTP studies and pesticide registrations), but no others. If the database lacks consistency, but there is reproducibility within the study, there should be no downgrading.
- g) The WG was not in agreement about the use of the word *causality* in the definitions. Some posited that the term *causality* is loaded in that independent experimental evidence can be necessary for proving causality. Others noted that determination of causality was an important issue in the NAS review of the EPA's IRIS assessment of formaldehyde. The NTP confirmed that while *causality* is not in the definitions, it is implied in the associations by incorporating the Hill Criteria.

Step 7: Integrate Evidence to Develop Hazard Identification Conclusions

Discussion Overview:

NTP presented a framework for integrating the evidence to develop hazard assessment conclusions that first combined the human and animal evidence, and then considered other relevant data. NTP indicates that other relevant evidence includes “ mechanistic data, *in vitro* data, and evidence on upstream indicators that might otherwise be overlooked.”

WG recommendations:

- 1) The WG suggests that the figures explaining the NTP approach (7A and 7B) should indicate that other relevant data could either increase or decrease the hazard ID conclusion (as presented they suggest it could only increase).
- 2) The WG suggests that evidence of no effect should be added to Figure 7A and B.
- 3) Upgrading (or downgrading) of conclusions should be done only with strong scientific evidence.

Specific WG comments for consideration by NTP:

- a) The table should make clear that when based solely on animal evidence, the hazard conclusion could be downgraded, in addition to be upgraded by using other relevant data (mechanistic data, secondary information, SAR data, metabolites, bioassays, endocrine assays).
- b) It should be possible for the NTP to upgrade a *presumed* to a *known* conclusion by using other relevant data (mechanistic data, secondary information, SAR data, metabolites, bioassays, endocrine assays).
- c) The WG suggests adding an additional conclusion, *known not to be a hazard*, when there is evidence of no health effect, while recognizing that this conclusion may be used infrequently.
- d) To avoid semantic difficulties, it needs to be made clear that it is possible to infer from limited direct evidence in humans that there is a hazard.

- e) Words like *known* can drive risk management decisions. The majority of the WG approved of using the word *known*, but it was recommended that one should not ignore the concerns regarding how this term may be subject to misinterpretation. Suggestions for words to use as a replacement for the word *known* were: *recognized*, *verified*, or *established*. The WG reached no agreement on this issue.
- f) The WG approved of the use of the same terms as the Globally Harmonized System (GHS); there is utility in consistency. However, it needs to be clear that the NTP system for using evidence is supported by the state-of-the-science. To avoid confusing end-users, the two processes should not be explicitly linked (i.e., NTP and GHS). On the other hand, an effort to align the language may facilitate the incorporation of new approaches that are innovated by the NTP into future updates of the GHS.
- g) One needs to consider how to use mutagenicity data and other short-term data. Do not discount high throughput assays at this point; their utility is context dependent and they may be very useful when linked to kinetic models.
- h) The WG members had differing opinions with respect to upgrading a conclusion of *inadequate* and *not classifiable* to *suspected* if data are available that provide strong scientific support. Supporting this idea is the anticipation of new types of scientific data (e.g., Tox21) with relevance to hazard assessment. Conversely, other WG members stated that the NTP methodology should be designed to account for these types of data earlier in the process. All WG members agreed that upgrading (or downgrading) of conclusions should be done only with strong evidence.
- i) With respect to using the word *known* as a category (Table 7b), concerns were noted to not ignore how this term may be subject to misinterpretation. Alternative suggestions as a replacement for the word *known* were: *recognized*, *verified*, or *established*. The WG did not reach an agreement on this issue.
- j) The WG listed types of other data that may be included: PK studies; *in vitro* – animal, human, outcome; wildlife; *in vivo* mechanistic/mode of action; transgenic studies; *in silico* – SAR; exposure assessment; nontraditional animal models (zebrafish and *C. elegans*).
- k) There are some difficult questions regarding incorporation of nontraditional animal studies, but they need to be considered.

- l) SAR can range from totally statistical to mechanistically hypothesized models.

Additional comments from the WG

Past NTP evaluations have been done differently for cancer and non-cancer endpoints by the Report on Carcinogens (RoC) and Center for the Evaluation of Reproductive Health Risks (CERHR), respectively. For these processes the NTP has always made the final hazard identification call, but the CERHR invoked more outside (nongovernmental) expert opinion prior to this step than the RoC. Some WG members expressed a preference for the CERHR process and others did not. However, the WG applauds the effort by NTP to produce a unified approach to the evaluation process.

The WG anticipates that the newly proposed process will be a major improvement over past practices in many respects. Past processes for both the RoC and CERHR have been models exemplifying high quality and transparent processes that have invoked public input at every stage resulting in public confidence. However, the advantage of the process presented to the WG will likely improve the reproducibility and transparency of assessments developed by NTP.

The WG suggests that NTP clarify the role of external expert scientific input in the evaluation process, as well as the role of contractors. Will the data extraction be done by NTP scientists or by contractors and what will be the qualifications for people who do data extraction? At what stage(s) will technical advisors be brought in to provide guidance (e.g., development of the protocol and identification of critical questions, review of evidence, development of conclusions)? The WG suggests that evaluations be conducted so that the necessary expertise is obtained in specific areas where it is needed (e.g., workshops). The process has been too much one-size-fits-all in the past.

The WG also notes that there has been an evolution in how NTP has engaged other federal agencies during the development of assessments. The WG suggests that the NTP clarify the process by which input from other federal agencies will be utilized throughout the development of an assessment. The process should be transparent and assurances secured that involvement from other federal agencies does not impede or delay completion of assessments by NTP.

The WG suggests that the NTP clearly explain how the public can track the assessment process and how they can provide meaningful input on how the evaluations are done.

APPENDIX A



National Toxicology Program

U.S. Department of Health and Human Services

NTP Board of Scientific Counselors Working Group on Reaching Evidence Assessment Conclusions

August 28 – 29, 2012

Lisa A. Bero, PhD
Professor
University of California, San Francisco
San Francisco, CA

Edward W. Carney, PhD
Senior Science Leader
Mammalian Toxicology
The Dow Chemical Company
Midland, MI

David C. Dorman, DVM, PhD
Associate Dean for Research and Graduate
Studies
College of Veterinary Medicine
North Carolina State University
Raleigh, NC

Elaine M. Faustman, PhD
Professor and Director
Institute for Risk Analysis and Risk
Communication
Department of Environmental and Occupational
Health Sciences
University of Washington
Seattle, WA

Lynn R. Goldman, MD, MPH
Dean and Professor
School of Public Health & Health Services
The George Washington University
Washington, DC

Dale Hattis, PhD
Research Professor
George Perkins Marsh Institute
Clark University
Worcester, MA

Malcolm Macleod, PhD
Reader and Head of Experimental Neuroscience
Centre for Clinical Brain
Sciences University of Edinburgh
Edinburgh, UK

Reeder L. Sams, II, PhD
National Center for Environmental
Assessment/RTP Division
US Environmental Protection Agency
Research Triangle Park, NC

Tracey J. Woodruff, MPH, PhD
Director, Program on Reproductive Health and
the Environment
University of California, San Francisco
Oakland, CA

Lauren Zeise, PhD
Chief
Reproductive and Cancer Hazard Assessment
Branch
Office of Environmental Health Hazard
Assessment
California Environmental Protection Agency
Oakland, CA



National Toxicology Program

U.S. Department of Health and Human Services

**DRAFT NTP APPROACH FOR REACHING CONCLUSIONS FOR
LITERATURE-BASED EVIDENCE ASSESSMENTS**

Division of the National Toxicology Program

National Institute of Environmental Health Sciences

National Institutes of Health

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

NTP Board of Scientific Counselors Working Group Meeting August 28-29, 2012

TABLES OF CONTENTS

INTRODUCTION	1
--------------	---

CONCEPTUAL OVERVIEW OF THE NTP FRAMEWORK FOR EVIDENCE ASSESSMENT	3
--	---

OBJECTIVES	3
------------	---

METHODS	4
---------	---

1.0	Prepare Topic	4
1.1	Develop draft statement of objectives and focus the preliminary question(s):	4
1.2	Prepare protocol for systematic review	5
1.3	Send out draft protocol to get input from appropriate experts and public	8
1.4	Revise statement of objectives and question(s) and protocol based on feedback	8
2.0	Search for and Select Studies for Inclusion	8
2.1	Perform systematic literature search based on revised strategy	9
2.2	Screen studies based on revised inclusion and exclusion criteria	9
3.0	Extract Data from Studies	9
4.0	Assess the Quality of Individual Studies	10
4.1	All Risk of Bias Elements	13
4.2	Major Risk of Bias Elements	14
5.0	Rate the Confidence in the Body of Evidence	16
5.1	Develop confidence rating within each study design type, for human and animal data separately:	17
5.2	If data are sufficient, consider performing meta-analysis and generating an estimate of effect within a study type (e.g., for all prospective studies)	26
5.3	Combine confidence conclusions for all study types	27
5.4	Develop confidence conclusions for multiple outcomes	27
6.0	Translate Confidence Ratings into Evidence of Health Effect	29
6.1	Evidence of Health Effects Descriptors	29
6.2	Consider evidence of effect or lack of effect	30
7.0	Integrate Evidence to Develop Hazard Identification Conclusions	32
7.1	Conclusions based on combination of evidence streams	32

7.2	Consider mechanistic, in vitro, or other supporting data	33
ACKNOWLEDGMENTS		34
REFERENCES		35

INTRODUCTION

The analysis program at the NTP conducts literature-based evaluations to assess the evidence that environmental chemicals, physical substances, or mixtures (collectively referred to as "substances") cause adverse health effects. Over the past year, the NTP has been working to utilize systematic review methodology in reaching conclusions for literature-based evidence assessments.¹ In brief, systematic review methods use a pre-specified approach to identify and critically appraise relevant research and to collect, report, and analyze data from the studies included in the review. The systematic review format helps provide a structure to guide identification and determination of literature for inclusion, as well as extraction of data from studies, assessment of study quality, and synthesis of data for reaching conclusions. Currently, systematic reviews are utilized most often in clinical epidemiology, although there is increasing interest in adopting this format in the environmental health sciences and risk assessment (EFSA 2010, National Research Council 2011, Woodruff and Sutton 2011, ATSDR 2012).

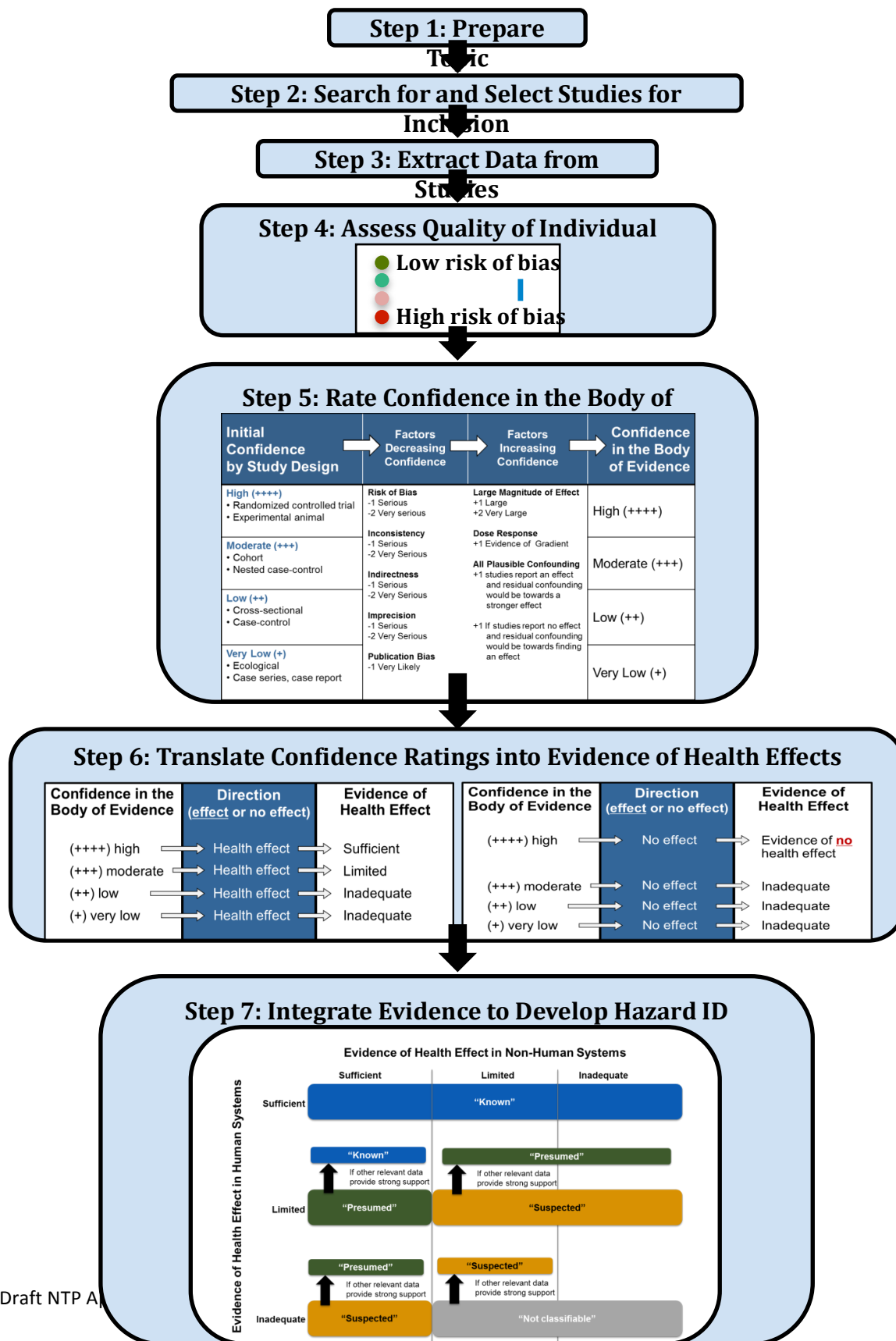
While adoption of systematic review will increase the transparency of how the NTP conducts literature-based evaluations, existing systematic review guidance is oriented toward health care (Guyatt *et al.* 2011a, Higgins and Green 2011, AHRQ 2012) and does not address the need within environmental health sciences to integrate across evidence streams (i.e., human, animal mechanistic data) in order to reach conclusions regarding potential health effects. There is some guidance on approaches to improve transparency and to assess data across a body of evidence to reach health care recommendations (Balshem *et al.* 2011, Guyatt *et al.* 2011a, Guyatt *et al.* 2012, Viswanathan *et al.* 2012); however, these approaches are based on evaluation of human studies often randomized clinical trials, and do not address many issues relevant to environmental health sciences. The NTP proposes an approach to integrate across evidence streams that is philosophically similar to the guidance established for health care but is tailored to the type of evidence considered in environmental health sciences.

The NTP's proposed methodology for reaching hazard identification conclusions for literature-based evidence assessments is outlined in this document in 7 steps (**Figure 2**). The NTP evidence assessment strategy builds on the rigorous and transparent evaluation of the data accomplished through systematic review. Evidence assessment refers to the process for reaching conclusions on a body of literature. For the NTP, this means assessing our confidence in the evidence within an evidence stream (i.e., human and animal data separately) and then integrating those conclusions across the evidence streams with consideration of other relevant data² such as supporting evidence from mechanistic studies.

¹ See <http://ntp.niehs.nih.gov/go/9741> for June 22, 2012 presentation to the NTP Board of Scientific Counselors and meeting minutes.

² See http://oehha.ca.gov/multimedia/green/pdf/GC_Regtext011912.pdf for definition and discussion of "Other relevant data"; in brief it refers to non-endpoint data, including chemical, physical, biochemical, biological or other data that may indicate a chemical substance may contribute to adverse effects.

Figure 1. NTP Framework for Conducting Literature-Based Evidence Assessment



CONCEPTUAL OVERVIEW OF THE NTP FRAMEWORK FOR EVIDENCE ASSESSMENT

A literature-based evidence assessment uses the information gathered by the systematic review process to reach hazard identification conclusions. The NTP's strategy for a literature-based evidence assessment ([Figure 2](#)) begins with a focused preparation of the topic and draft protocol (step 1). The protocol documents the specific approach to be used in the evaluation including key aspects of searching for and selecting studies, grouping related outcomes, extracting data, and evaluating "study quality" or risk of bias of individual studies. Steps 2 and 3 then follow the process outlined in the protocol to search for studies, select studies for inclusion, and extract data from the selected studies. Steps 4 and 5 begin to evaluate the quality of the evidence. First, in step 4, "study quality" or risk of bias of individual studies is assessed for specific outcomes by using a set of questions to evaluate study design and performance (see [Table 1](#)). Then, the confidence in the body of evidence across studies for specific outcomes is assessed based on study designs, limitations, and strengths in step 5 (see [Table 3](#)). The confidence ratings are integrated across outcomes within an evidence stream (human or non-human), ideally based on biologically related groups of outcomes specified a priori in the protocol. Then in step 6, these confidence ratings are translated into an evidence of health effects category of "sufficient", "limited", or "inadequate" (see [Table 4](#) for definitions) by considering the direction of the effect (toxicity or no toxicity) (see [Table 6A](#) and [Table 6B](#)). In general, the assessments conducted by the NTP will be qualitative as quantitative analysis (i.e. meta-analysis) will not be reasonable for datasets with a wide variety of environmental health data contributing to conclusions. In step 7, the NTP then integrates the evidence of health effects determinations for the human and animal evidence streams to reach an overall hazard identification classification of "known," "presumed," "suspected," or "not classifiable" (see [Table 7A](#) and [Table 7B](#)). Other relevant data such as supporting evidence from mechanistic studies is also considered when reaching hazard identification conclusions.

The NTP has considered efforts underway both internationally and within the United States toward use of the Globally Harmonized System for Classification and Labeling of Chemicals (GHS) for identification of toxicants. Consistent with GHS, the NTP is adopting the following categories for hazard identification³:

- Known to be a hazard to humans [equivalent to Category 1A GHS]
- Presumed to be a hazard to humans [equivalent to Category 1B GHS]
- Suspected to be a hazard to humans, [equivalent to Category 2 GHS]
- Not classifiable or not identified to be a hazard to humans

The process laid out in this document allows for a rigorous evaluation of diverse sets of information to reach hazard identification conclusions for the overall body of evidence.

OBJECTIVES

These methods are designed to foster a transparent and consistent systematic approach for evidence assessment in reaching and communicating conclusions in NTP literature-based evaluations. In order

³ Currently a different set of congressionally hazard identification categories are used for listing substances in the NTP *Report on Carcinogens*. The categories in this table would be used for evaluations other than for that report. [congressionally mandated categories for identification of cancer hazards]

to identify potential environmental hazards to human health, the NTP supports using a health protective approach when scientific judgments are made.

METHODS

The 7 steps outlined in this section provide guidance, and application of the methods will involve scientific judgments on a case by case basis. However, the approach outlined below encourage transparency in documenting the basis for scientific decisions, and this strength is a key aspect of the NTP approach for reaching hazard identification conclusions for literature-based evidence assessments.

Steps in the draft NTP approach for reaching conclusions for literature-based evidence assessments

1. Prepare topic
2. Search for and select studies for inclusion
3. Extract data from studies
4. Assess quality of individual studies
5. Rate the confidence in the body of evidence
6. Translate confidence into evidence of health effects
7. Integrate evidence to develop hazard identification conclusions

4.0 PREPARE TOPIC

Preliminary objectives and questions should reflect informed judgment about the ability of the literature base to answer specific questions and/or identify areas of uncertainty and data gaps. These questions should specify the populations, exposures, and outcomes that are of interest which are then used to develop *a priori* eligibility criteria for inclusion of studies in the review. Ideally questions and objectives should be formed before initiating the full review, but they should not prevent the exploration of unanticipated issues that might arise (Khan *et al.* 2001). At this stage, the objectives and questions begin to focus the systematic review and facilitate expert input; however, it is critical to document the refined objectives and questions in the draft protocol and any modifications made during completion of the systematic review.

4.1 Develop draft statement of objectives and focus the preliminary question(s): Consult with appropriate experts to focus question(s) based on the need and availability of data.

4.1.1 The Objectives could be either to

- Answer a specific question or questions, or
- Use the systematic review to identify areas of uncertainty, data gaps, etc.

4.1.2 Objectives should consider PECO or PICO principles:

- **Population of interest:** specify characteristics of study participants in such way as to not force the unnecessary exclusion of studies based on recent or expensive diagnostic criteria or the type of setting of the study
- **Exposure or Intervention:** common or core features of the exposure should be specified when the exposure or intervention is complex

- **Control or comparator** group: specify groups against which the exposure will be compared based on the goals of the evaluation
- **Outcomes** of interest: pre-specified as primary (clear human health effects), and secondary (surrogate measures, not clinical endpoints)

4.1.3 A preliminary search and consultation with librarian early in the process may help with focusing objectives and questions.

4.2 Prepare protocol for systematic review

4.2.1 Develop draft protocol

4.2.1.1 Systematic Literature Search

A thorough and reproducible search of a range of sources is required to identify as many relevant studies as possible. Bias in identifying studies can be minimized by searching multiple databases and avoiding restrictions on publication type or language. In cases where the question to be addressed is very broad, exploratory screening of the literature can be useful to help refine the focus of an evaluation.

Specific Recommendations:

- Conduct comprehensive literature search.
- Collaboration with librarian trained in systematic review methods is **strongly** recommended, and if this is not done, considerable external review of the literature search methods should be obtained.
- Search multiple databases (including but not limited to PubMed, TOXNET, EMBASE, Scopus, Web of Science, Cochrane Library, etc.).
- Detail the search strategy in the protocol such that it could be reproduced for at least one database.
- Provide details on the date of the search, whether it will be updated and when, and any limits placed on the search (such as by language or date of publication).
- Document additional sources of references such as those identified by reference sections of review articles or by experts in the field.
- Establish minimum requirements for inclusion of data from meeting abstracts, other unpublished “gray literature,”⁴ or author communications.

4.2.1.2 Selection of Studies

Identifying all studies relevant to a question from a comprehensive literature search requires clear and consistent guidelines. Determining which studies to include is one of the most influential decisions made in the review process (Higgins and Green 2011). Using multiple independent reviewers can help ensure that judgments made in the process are

⁴ literature that is not formally published in sources such as books or journal articles, including conference abstracts
Draft NTP Approach for Consideration at August 28-29, 2012 BSC Working Group Meeting

documented and reproducible. The exact approach used for each review will vary, but the protocol should include those details.

4.2.1.3 Establish specific plan for reviewing studies for inclusion (typically 2 reviewers reviewing every reference)

4.2.1.4 Establish the approach for resolving conflicts between reviewers (e.g., joint review for decision or 3rd reviewer to make decisions)

4.2.1.5 Establish criteria for inclusion and exclusion of references pertinent to the questions and document the reasoning for these criteria based on

- Outcomes of interest
- Relevant exposures
- Types of studies or outcomes pre-specified as not pertinent

4.2.2 Determine grouping and hierarchy of outcomes

Related outcomes may be considered together or separately depending on their clinical relevance, similarity of mechanisms, or directness in representing human health effects. The protocol should include as much detail as possible to specify which outcomes will be considered jointly and if they will be considered primary (health effects) or secondary (surrogate measures or upstream indicators) outcomes. These decisions will inevitably be adjusted during completion of the review based on the available data, but a distinction between *priori* and *pos hoc* decisions should be documented and justified in the protocol.

Specific Recommendations:

- Group outcomes based on relatedness of the outcome measures (e.g., HbA1C and high fasting glucose as diagnostics for diabetes).
- Document reasoning and support for grouping or splitting of outcomes
- Establish which outcomes will be considered primary (health effects) or secondary (surrogate measures)

4.2.3 Develop project-specific forms for data extraction

The NTP has developed a data extraction forms to capture study attributes important for the evaluation (e.g., exposure specifics, timing, outcomes, geographic location for human studies, species and strain for animal studies, etc.) in a systematic manner. Separate template forms have been developed for human, animal, in vitro, and meta-analysis studies (See Appendices). These forms will have customized content for each review and will be included in the protocol. To the extent possible, the forms utilize controlled vocabulary to minimize variation between individual data extractors. This controlled vocabulary can be expanded or revised during the completion of the systematic review as needed for a given project.

Specific Recommendations:

- *Develop a controlled vocabulary for categories of health effects and specific outcomes under investigation.*
- *Develop a controlled vocabulary for exposure assessment methods likely to be used and reported in studies.*
- *Identify critical cofactors to be identified in adjusted analyses.*
- *Establish specific plan for extracting data (typically extractors, or single extractor and quality control system).*
- *Establish the approach for quality control (between data extractors, or quality control system).*
- *Add additional project-specific questions necessary to fully evaluate study results.*

4.2.4 Establish a plan for how risk of bias for individual studies will be evaluated

Risk of bias of individual studies, or internal validity, refers to “the extent to which a single study’s design and conduct protect against all bias in the estimate of effect” (Viswanathan *et al.* 2012). The term “study quality” has been used to pertain to many aspects of risk of bias as well as overall confidence in the body of evidence, so wider use of the term “risk of bias” is encouraged to avoid ambiguity. The protocol should specify project-specific details of how risk of bias will be evaluated for each exposure and outcome under review.

Complete details of how this method will be implemented are included in section **7.0 Assess the Quality of Individual Studies**. The NTP’s specific risk of bias questions are detailed in **Table 1**. The questions in **Table 1** are geared to address risk of bias in studies that can be used to evaluate the potential adverse health effects of environmental chemicals, physical substances, or mixtures. A subset of 4 “major” risk of bias questions is also identified in section **7.2 Major Risk of Bias Elements**. These questions address risk of bias issues that may have a greater impact on confidence in data for drawing conclusions on health effects of environmental chemicals.

Specific Recommendations:

- *Establish specific plan for reviewing studies for risk of bias (typically 2 independent evaluators or 1 evaluator with review).*
- *Outline how inconsistencies between risk of bias reviewers will be handled*
 - *Two reviewers can make joint review decision or bring in third reviewer*
 - *If the approach is to use one evaluator with independent review, outline the plan by which differences in judgment will be resolved between the original evaluator and the reviewing team member.*
- *Outline and justify research question-specific definitions for what constitutes differing levels of risk of bias from low to high risk of bias.*
- *In most reviews, studies that do not report details pertaining to a risk of bias question will be rated as “probably a high risk of bias.” Explain and justify if studies with poor reporting will be handled differently.*

- *If applicable to the objectives, add additional risk of bias questions under the “Other” category and justify the rationale for the additional questions.*
- *The methods identify major risk of bias elements in [section 7.2](#) Explain and justify if the reviewers would consider a different set of questions as major elements based on the objectives of the evaluation.*
- *As necessary, modify project-specific risk of bias forms.*

4.3 Send out draft protocol to get input from appropriate experts and public

Development of a systematic review protocol is best accomplished in consultation with appropriate subject matter experts and is typically an iterative process. To obtain external input, the NTP, for example, might engage technical advisors or solicit public comments on the draft protocol. Detailed preparation of the protocol will allow many scientific judgments to be made *a priori* and reduce the number of revisions made during completion of the review. This approach will increase scientific input and transparency of the review process. The following issues are potential topics for outside scientific input.

Specific Recommendations:

- **Question(s) to be answered by the systematic review:** *ask if the questions are appropriately focused, and if not, how could they be changed so that the question is neither too focused and thereby risks missing an important related issue, or too broad to allow reasonably sized database focusing of question; also ask for suggestions of potential prioritization among sub-questions if there are more than one question*
- **Identification and definition of outcomes** *ask if the list of outcomes is appropriate, or if the list should be increased or decreased, and if so, in what way; also ask for comments on the specific wording and definitions of the outcomes.*
- **Grouping of outcomes:** *ask for feedback on the proposed grouping of outcomes based on related mechanisms or mode of action.*
- **Identification of outcomes as primary (health effects) or secondary (surrogate measures or upstream indicators)** *for protecting human health.*
- **Identification of outcomes as “not relevant” to the question**
- **Refine search strategy:** *ask for suggestions to focus or further limit the search along with comments on completeness to answer the questions at hand*
- **Feedback on risk of bias questions and which questions are “major” elements**
- **Identification of problematic (a) outcome assessment, (b) exposure assessment, or (c) other data issues** *ask experts for this subject matter if they can identify a priori issues (including higher risk of bias) that would limit the ability of study to answer the questions specific to this review; also ask reviewers to identify if any of the above issues are so damaging to the confidence that resulting data may have little to no utility for drawing conclusions.*

4.4 Revise statement of objectives and question(s) and protocol based on feedback

5.0 SEARCH FOR AND SELECT STUDIES FOR INCLUSION

While this step and the following ([6.0 Extract Data from Studies](#)) are the most labor-intensive in the process, if the protocol is sufficiently detailed, few scientific judgments will need to be made

during these steps. Unanticipated issues will undoubtedly arise, requiring revisions to the protocol, which will be documented.

5.1 Perform systematic literature search based on revised strategy

- *Document the date of the literature search and any additional searches done to update the available data based on decisions made in the protocol.*
- *Remove duplicate references.*

5.2 Screen studies based on revised inclusion and exclusion criteria

A comprehensive literature search will identify a large number of references that will need to be screened in a systematic and transparent manner to identify the information relevant to the question as detailed in the protocol.

Additional sources of information, not identified in the literature search can be included as well. In situations where relevant data have not undergone independent, external peer review, the NTP will ensure they are appropriately peer reviewed prior to their inclusion in an assessment. All information included will be publicly available.

Specific Recommendations:

- *Use 2 independent reviewers for screening (review every study in duplicate).*
- *Implement plan for resolving conflicts between reviewers as stated in the protocol (e.g., joint review and decision, or 3rd reviewer).*
- *Follow designated protocol to proceed by reviewing first the title, then the abstract, and finally the full text, or combined title and abstract screening, etc.*
- *References with uncertain relevance should remain under consideration until the full text screening stage.*
- *Document the progress of screening studies through this process and construct a flow diagram outlining the number of articles included or excluded at each stage and the reasons for exclusion.*
- *Initial “training” set of studies should be pre-screened to help the reviewers establish a consistent approach to screening studies.*
 - *The criteria established in [step 4.2.1.5](#) should be applied.*
 - *set of studies is suggested for training with a range of study types to test various aspects of the screening criteria.*
 - *Any adjustments to the screening criteria should be documented and will eventually be included in a revised protocol to be published with the results of the evaluation.*

6.0 EXTRACT DATA FROM STUDIES

See Appendices for template data extraction forms for humans, animals, and meta-analysis studies.

Specific Recommendations:

- *Data should be extracted by two data extractors and reconciled or extracted by one data extractor and verified by reviewer through documented quality control procedures.*

- *The NTP is still considering the most efficient approach for implementing quality control measures for data extraction. The minimal quality control procedures implemented should include duplicate independent data extraction on the principal studies used to develop conclusions.*
- *An initial “training” set of studies should be extracted to help the data extractor establish a consistent approach.*
 - *The data extraction forms established in [step 4.2.3](#) should be utilized.*
 - *set of studies with a range of study types will be used to test various aspects of the extraction forms and for training data extractors.*
 - *Any adjustments to the data extraction forms should be documented and will eventually be included in a revised protocol to be published with the results of the evaluation.*
- *Grouping decisions and determination of primary (health effect) and secondary (surrogate measures or upstream indicators) outcomes should be considered in extracting the data so that related outcomes can be easily evaluated together or separately.*
- *Data extracted in developing NTP reviews will be made publically available (e.g., they might be stored in the NTP Chemical Effects in Biological Systems (CEBS) database) for data mining after completion of the review.*

7.0 ASSESS THE QUALITY OF INDIVIDUAL STUDIES

Risk of bias is a preferred term by systematic review methodologists to encompass several aspects of internal validity. This term is not synonymous with “study quality,” which has been used as a more general term that includes many aspects of risk of bias in addition to overall confidence in the body of evidence or ability of study results to contribute to a health effect conclusion.

There is no existing consensus for how to assess risk of bias for observational human studies or experimental animal studies in systematic reviews. The NTP approach is based on guidance from “Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions” in the Methods Guide for Comparative Effectiveness Reviews from the Agency for Healthcare Research and Quality (AHRQ) (Viswanathan *et al.* 2012). [Table 1](#) includes 18 specific questions within five domains (selection, performance, attrition, detection, and reporting bias) and outlines which types of study designs they are applicable to (randomized controlled trials, cohorts, nested case-control studies, cross sectional studies, case-control studies, case series, case reports, and experimental animal studies).

Risk of bias criteria for animal studies are not addressed by clinical epidemiology guidelines. NTP will apply criteria to experimental animal studies that are similar to criteria applied to human randomized controlled trials because these study designs are similar in ability to control timing and dose of exposure and minimize confounding factors. Using the same set of questions for all study types, including experimental animal studies, allows for comparison of particular risk of bias issues with a common language across all studies.

Table 1: NTP Risk of Bias Questions

Types of Bias	Design-specific Risk of Bias Questions*	Experimental	Animal	RCT	Cohort	Case-control	Cross-sectional	Case Series
		I						
Selection	Was treatment or exposure adequately randomized?	x		x				
	Was treatment or exposure allocation adequately concealed?	x		x				
	Was the subject recruitment strategy uniform across study groups?			x	x			
	Is the comparison group appropriate, including similar baseline characteristics and having the exposed and non-exposed subjects drawn from the same population?	x		x	x	x	x	x
	Does the study design adjust/control for important confounding and modifying variables?	x		x	x	x	x	x
Performance	Did researchers adjust/control for other exposures or interventions that may bias results?	x		x	x	x	x	x
Attrition	In RCT, animal, or cohort studies: does the length of follow-up differ between groups? In case-control studies: is the time period between exposure/intervention and outcome the same for cases and controls?	x		x	x	x		
	Was the attrition rate uniformly low?	x		x	x	x		
	Is the analysis conducted on an intention-to-treat basis?			x	x			
	Was follow-up long enough to assess the outcome of interest?	x		x	x			
Detection	Can we be confident that the outcome did not precede exposure?	x		x	x			x
	Were outcome assessors blinded to the exposure or intervention status of participants?	x		x	x	x	x	x
	Is inclusion/exclusion criteria measured reliably, implemented consistently?	x		x	x	x	x	x
	Are confounding variables assessed using reliable measures, implemented consistently?				x	x	x	x
	Are data analyses appropriate, performed with reliable tests, implemented consistently?	x		x	x	x	x	x
	Can we be confident in the exposure assessment?	x		x	x	x	x	x
	Can we be confident in the outcome assessment?	x		x	x	x	x	x
Reporting	Are outcomes pre-specified by the researchers? Are all pre-specified outcomes reported?	x		x	x	x	x	x
Other								

RCT, randomized controlled trials;





An “x” at the intersection of a question-row and study-type-column indicates the risk of bias question applies to that study type.

This table is adapted from the AHRQ publication “Assessing the Risk of Bias of Individual Studies when Comparing Medical Interventions” (Viswanathan *et al.* 2012).

*NOTE: As risk of bias increases, confidence in a study's results decreases. All risk of bias questions have been consistently worded such that an answer of "Yes" corresponds to positive study attributes and therefore a low risk of bias. Keep in mind that in order to keep the wording straight forward and consistent across studies, the questions are not worded to answer the question "Is there risk of bias?"

7.1 All Risk of Bias Elements

Specific Recommendations:

- Risk of bias is evaluated for individual studies on a outcome specific basis. Certain aspects of study design and conduct may increase risk of bias for some outcomes and not others.
- Each of the 18 questions is answered on a 4 point scale (Guyatt 2012).
 - **Definitely Yes** Definitely Low risk of bias.... 
 - **Probably Yes** Probably Low risk of bias..... 
 - **Probably No** Probably High risk of bias..... 
 - **Definitely No** Definitely High risk of bias.... 
- All 18 questions are not applicable to all study designs. [Table 1](#) identifies questions as relevant for a particular study design with an “x” in [Table 1](#) at the intersection of the study question row and the study type column.
- If a study fails to report sufficient information to evaluate the risk of bias question, “Probably High risk of bias” will be the default answer unless otherwise specified and justified in the protocol. If the information is not explicitly reported but can be inferred then “probably yes” or “probably no” is used as risk of bias response.
- Project-specific criteria for what constitutes high or low risk of bias are defined in the protocol.
See [Example Box 4.1A](#).
- Additional risk of bias questions to cover a risk of bias that is not already covered in the default 18 questions can be added and specified in the protocol.
- An initial “training” set of studies should be evaluated for risk of bias to help reviewers establish a consistent approach.
 - The plan established in [step 4.2.4](#) the questions as stated, and the “major” risk of bias elements should be utilized.
 - Use of a set of studies with a range of risk of bias issues is suggested for training to test various aspects of answering the risk of bias questions and the operational definitions documented in [step 4.2.4](#).
 - Any adjustments to the operational definitions, use of other questions, or designation of major risk of bias questions should be documented and included in a revised protocol that is published with the results of the evaluation.
- All risk of bias questions should be considered when determining confidence in the body of evidence ([step 8.0](#)).

Example Box 4.1A: Risk of bias

PROJECT-SPECIFIC CRITERIA THAT DEFINE RISK OF BIAS LEVELS BASED ON STUDY ATTRIBUTES

Risk of bias question:

Can we be confident in the outcome assessment of diabetes?

Definitions:

- **Definitely Low risk of bias:** Self-report of diabetes medication, hospital records, clinically accepted diagnostic criteria
- **Probably Low risk of bias:** self-report diabetes or doctor’s diagnosis
- **Probably High risk of bias:** death certificate with a large sample size
- **Definitely High risk of bias:** death certificate

- In summary tables or figures of different study types, the use of fifth category of “Not Applicable” or “NA” is suggested to indicate some questions are not appropriate for some studies based on study design.
- See [Example Box 4.1B](#) for an example of displaying risk of bias criteria across multiple studies.

Example Box 4.1B: Displaying risk of bias across studies

RISK OF BIAS RATINGS CAN BE DISPLAYED ACROSS MULTIPLE STUDIES TO AID IN EXAMINING POTENTIAL BIAS WITHIN AND ACROSS STUDIES

Types of Bias	Risk of Bias Questions	Andy et al., 2010	Bob et al., 1999	Carly et al., 2000	David et al., 2011	Evita et al., 2008
Selection	Was treatment or exposure adequately randomized?	●	○	○	○	○
	Was treatment or exposure allocation adequately concealed?	●	○	○	○	○
	Was the subject recruitment strategy uniform across study groups?	○	●	●	○	○
	Is the comparison group appropriate, including similar baseline characteristics and having the exposed and non-exposed subjects drawn from the same population?	●	●	●	●	●
	Does the study design adjust/control for important confounding and modifying variables?	●	●	●	●	●
Performance	Did researchers adjust/control for other exposures or interventions that may bias results?	●	●	●	●	●
Attrition	In RCT, animal, or cohort studies: does the length of follow-up differ between groups?	●	●	●	○	○
	In case-control studies: is the time period between exposure/intervention and outcome the same for cases and controls?	●	●	●	○	○
	Was the attrition rate uniformly low?	●	●	●	○	○
	Is the analysis conducted on an intention-to-treat basis?	○	●	●	○	○
	Was follow-up long enough to assess the outcome of interest?	●	●	●	○	○
Detection	Can we be confident that the outcome did not precede exposure?	●	●	●	○	○
	Were outcome assessors blinded to the exposure or intervention status of participants?	●	●	●	●	●
	Is inclusion/exclusion criteria measured reliably, implemented consistently?	●	●	●	●	●
	Are confounding variables assessed using reliable measures, implemented consistently?	○	●	●	●	●
	Are data analyses appropriate, performed with reliable tests, implemented consistently?	●	●	●	●	●
	Can we be confident in the exposure assessment?	●	●	●	●	●
Reporting	Can we be confident in the outcome assessment?	●	●	●	●	●
	Are outcomes pre-specified by the researchers? Are all pre-specified outcomes reported?	●	●	●	●	●

Note: White circles for NA or does not apply are filled out for the hypothetical studies as follows based on study design:


- Andy et al., 2010 is an experimental animal study
- Bob et al., 1999 and Carly et al., 2000 are cohort studies
- David et al., 2011 and Erica et al., 2008 are cross-sectional studies

7.2 Major Risk of Bias Elements

The NTP will designate a subset of risk of bias questions as major criteria critical to the evaluation of environmental health studies. While many aspects of study design are important to assessing potential bias, these risk of bias elements have a greater impact on confidence in the evidence that environmental substances are associated with health effects. There are 4 default major risk of bias elements questions ([Table 2](#)) and they are considered major factors for study types as follows.

Table 2: Major areas of risk of bias for environmental health questions by study type

Major Risk of Bias Questions	Experimental Animal	Human Studies				
		RCT	Cohort	Case control	Cross sectional	Case Series
Can we be confident in the exposure assessment?	x	x	x	x	x	x
Can we be confident in the outcome assessment?	x	x	x	x	x	x
Is the comparison group appropriate, including similar baseline characteristics and having the exposed and non-exposed subjects drawn from the same population?	x	x	x	x	x	x
Does the study design adjust/control for important confounding and modifying variables?	x	x	x	x	x	x

An “x” at the intersection of a question-row and study-type-column indicates the risk of bias question applies to that study type. A shaded box  is used to identify that the question is considered a “major” question for that study type.

Specific Recommendations:

- **Can we be confident in the exposure assessment?**
 - Exposure assessment is considered a major risk of bias factor for all study types (experimental animal and human RCT, cohort, case-control, cross-sectional studies, and case series).
 - Exposure assessment is particularly important for human studies relevant to environmental health questions because human studies are unlikely to include intentional exposure studies because of ethical considerations in exposing humans to suspected environmental toxicants.
 - Appropriate and reliable chemical characterization is required to be confident that the observed effect (or lack of effect) is associated with exposure to the substance in question.
 - For observational human studies, exposure data can be minimal and imprecise such as stationary workplace monitors rather than personal monitors, or there are few data points covering minimal time period; therefore, exposure assessment is considered a major risk of bias for environmental health questions.
- **Can we be confident in the outcome assessment?**
 - Outcome assessment is considered a major risk of bias factor for all study types (experimental animal and human RCT, cohort, case-control, case series, and cross-sectional studies).
 - Accurate and consistent outcome assessment is important for comparability across studies and to avoid differential misclassification.

- ***Is the comparison group appropriate, including similar baseline characteristics and having the exposed and non-exposed subjects drawn from the same population?***
 - *The use of a appropriate comparison group is important for all study types.*
 - *For observational human studies, selection of appropriate comparison group is particularly important, because workplace or geographic regions are often used as the basis on which to define the exposed group. Ideally, the comparison group should be from the same region or workplace. Selection of the comparison group is considered a major risk of bias for environmental health questions.*
 - *For experimental studies, the use of an appropriate vehicle-exposed control group is critical and should be considered here.*
 - *For experimental studies, age, sex, weight, and adequate sample size should be similar between groups.*
- ***Does the study design adjust or control for important confounding and modifying variables?***
 - *Adjustment and confounding are critical for observational studies and the human data to answer environmental health questions are likely to be restricted to observational studies.*
 - *For experimental studies, confounding may occur; however experimental studies can limit or eliminate confounding through randomization of treatment, allocation to groups, etc. Therefore, confounding and modifying variables should be considered for experimental animal and RCT studies, but these risk of bias areas are unlikely to be “major” factors for experimental studies.*
- *Reviewers are encouraged to not exclude studies that could inform the overall conclusions of the evaluation; however, information gained from studies with ratings of definitely high in 2 or more major risk of bias elements is unlikely to be informative because of the significant risk of bias determined in those studies.*
 - *Thus, if a study is rated “Definitely High” risk of bias for two or more major risk of bias elements on a given outcome, results will be excluded and not be considered in developing conclusions for that outcome.*
 - *It is possible that studies that have been excluded due to ratings of Definitely High in 2 or more major risk of bias elements could be considered in sensitivity analyses at the completion of the evaluation.*

8.0 RATE THE CONFIDENCE IN THE BODY OF EVIDENCE

Confidence in the body of evidence, or confidence that the association or the estimates of effect are correct, is assessed by considering the strengths and weaknesses of a group of studies of similar study design. The NTP method developed for environmental health hazard identification is based on the GRADE⁵ (which has been adopted by the Cochrane Collaboration (Schünemann *et al.* 2012)) and AHRQ approaches, which are conceptually very similar (Balslem *et al.* 2011, Lohr 2012). For each outcome, groups of studies are given an initial confidence rating by study design. Evaluators should then develop a

⁵ The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (<http://www.gradeworkinggroup.org/>)

confidence rating for the body of evidence for each study design (e.g., for prospective cohort studies separately from cross-sectional studies). The initial rating is downgraded

Table 3: Schematic to Develop Confidence Rating for the Body of Evidence

Initial Confidence by Study Design	Factors Decreasing Confidence	Factors Increasing Confidence	Confidence in the Body of Evidence
High (++++) <ul style="list-style-type: none"> • Randomized controlled trial • Experimental animal 	Risk of Bias -1 Serious -2 Very serious	Large Magnitude of Effect +1 Large +2 Very Large	High (++++)
Moderate (+++) <ul style="list-style-type: none"> • Cohort • Nested case-control 	Inconsistency -1 Serious -2 Very Serious Indirectness -1 Serious -2 Very Serious	Dose Response +1 Evidence of Gradient All Plausible Confounding +1 studies report an effect and residual confounding would be towards a stronger effect	Moderate (+++)
Low (++) <ul style="list-style-type: none"> • Cross-sectional • Case-control 	Imprecision -1 Serious -2 Very Serious	+1 If studies report no effect and residual confounding would be towards finding an effect	Low (++)
Very Low (+) <ul style="list-style-type: none"> • Ecological • Case series, case report 	Publication Bias -1 Very Likely		Very Low (+)

for factors that decrease confidence and upgraded for factors that increase confidence in the results. Then, confidence across all available study designs is assessed. A single, well conducted study may provide evidence of toxicity or a health effect associated with exposure to the substance in question (e.g., see Germolec (2009) and Foster (2009) for explanation of the NTP levels of evidence for determination of “toxicity” for individual studies). If a sufficient body of very similar studies is available, a quantitative meta-analysis can generate an overall estimate of effect. Finally, confidence conclusions are developed across multiple outcomes considering biologically-related and unrelated outcomes. The NTP recognizes that the scientific judgments involved in these ratings are inherently subjective, but the process outlined here (Table 3) provides a transparent framework to document and justify decisions made in arriving at a confidence rating.

8.1 Develop confidence rating within each study design type, for human and animal data separately:

8.1.1 Definitions for confidence ratings for outcomes

The four ratings noted above can be applied to describe the confidence in a body of evidence. The approach outlined in steps 8.1.2–8.1.4 describes important factors to consider and the process to follow in developing confidence conclusions for the body of evidence. These definitions are conceptually similar to those used by GRADE, the Cochrane Collaboration,

and AHRQ to describe four levels of “quality” of evidence for quantitative meta-analyses (Balshem *et al.* 2011, Lohr 2012), but they have been modified so they are better suited to evaluations that are more qualitative in nature (i.e. “apparent relationship rather than “estimate of effect”). In the context of identifying research needs, a conclusion of “high confidence” could also be interpreted as further research is very unlikely to change our confidence. Conversely, a conclusion “low confidence” would suggest that further research is very likely to have an important impact on our confidence.

- **High Confidence (++++)** in the association between exposure to the substance and the outcome. The true effect is highly likely to be reflected by the apparent relationship.
- **Moderate Confidence (+++)** in the association between exposure to the substance and the outcome. The true effect may be reflected in the apparent relationship.
- **Low Confidence (++)** in the association between exposure to the substance and the outcome. The true effect may not be reflected in the apparent relationship.
- **Very low Confidence (+)** in the association between exposure to the substance and the outcome. The true effect is highly likely to be different than the apparent relationship.

8.1.2 Initial confidence set by study design for each outcome

Experimental studies (animal or human RCT) are by design better suited to address causality due to temporality or confidence that exposure preceded outcome and this is reflected in the highest initial confidence rating. Experimental animal studies are similar in design to human RCTs and both will have an initial confidence rating of “high.” For environmental health questions, randomized controlled trials (RCTs) will rarely be available, and observational studies will comprise the bulk of the human data on the substance in question. Observational studies vary significantly in their ability to inform the true association between exposure to a substance and the health effect. To distinguish between the overall strengths and weaknesses of these study designs, observational studies are stratified into three different initial confidence levels (“moderate”, “low”, or “very low”).

Cohort studies and case-control studies nested within cohort studies follow participants over time and can provide longitudinal information about the timing of exposure and changes in outcomes or the development of disease. Cross-sectional studies provide information on exposures and outcomes at the same point in time, so it is possible that the outcome of interest was present before the exposure or even contributed to the observed exposure level. Case-control studies compare exposure levels in disease cases to disease-free controls to estimate the association; but selection of proper controls is critical, retrospectively collected exposure information is prone to

error, and the exposures may not have preceded the development of disease. Ecological studies look at exposure and outcomes on a population level and do not have information on individual participants. Such correlations on the population level are not appropriate to apply to individual level estimates of risk. Case reports and case series have small sample sizes and no comparison groups, so they cannot reliably estimate the association between an exposure and an outcome. Both ecological studies and case series studies can generate hypotheses for subsequent studies with stronger study designs, but there is very low confidence to base conclusions on only these types of studies.

It is important to emphasize that study design is not a proxy for determining the confidence in individual studies. Rather, it is a starting point that reflects features of each study type, but studies will be evaluated individually.

Initial Confidence Rating by Study Design

- **High Confidence (++++)**
 - *Randomized controlled trial studies*
 - *Experimental animal studies*
- **Moderate Confidence (+++)**
 - *Prospective*
 - *Nested case-control*
- **Low Confidence (++)**
 - *Cross-sectional studies*
 - *Case-control studies*
- **Very low Confidence (+)**
 - *Ecological studies*
 - *Case series or case study*

8.1.3 Downgrade confidence rating

Through a collaborative effort spanning more than a decade, GRADE and AHRQ have identified five categories for downgrading confidence in a body of evidence that captures the main issues (risk of bias, inconsistency, indirectness, imprecision, and publication bias) that could decrease confidence in a body of evidence (Balslem *et al.* 2011, Lohr 2012). The factors that can downgrade confidence are discussed in more detail below in sections [8.1.3.1](#) - [8.1.3.5](#).

In reality, confidence is a continuum that has been somewhat artificially categorized in this system. As such, the reasons for downgrading confidence may not fit neatly into a single category. If the decision to downgrade is borderline for two categories, the body of evidence can be downgraded one level for one of the two categories to account for both partial concerns. In

addition, the body of evidence should not be downgraded twice for what is essentially the same limitation that could be considered applicable to more than one category.

8.1.3.1 Risk of bias of the body of evidence

Risk of bias criteria have been described previously in section [7.0 Assess the Quality of Individual Studies](#) where study quality issues for individual studies are evaluated on an outcome specific basis. In this step, the overall risk of bias for the entire body of evidence is considered. Again, each outcome should be considered separately because the source of bias may vary in importance between outcomes. If there are areas with high risk of bias that lower confidence in the overall conclusions, the rating can be downgraded one or two levels (Guyatt *et al.* 2011g).

Assessing risk of bias across studies is a key component of evaluating confidence in the body of evidence; however, identifying risk of bias limitations can also be instrumental for informing future research to target those areas that will most directly increase confidence in the body of evidence.

Specific Recommendations:

- *Use of “major” risk of bias criteria to exclude studies with high risk of bias will increase confidence in the overall body of evidence on which conclusions are based.*
- *Consider the entire body of evidence for all risk of bias criteria.*
- *Downgrade one or two levels for an overall decrease in confidence.*

8.1.3.2 Inconsistency

Inconsistency, or large variability in the magnitude of estimates of effect, reduces confidence in the body of evidence.⁶ Inconsistency in terms of the direction of effect should also be considered here.

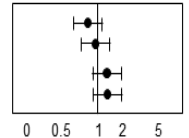
Large inconsistency across a body of literature should be further explored, preferably through *a priori* hypotheses that might explain the heterogeneity. If there is less inconsistency within subgroups of the body of evidence (e.g., men versus women), the protocol can be amended to consider these groupings separately.

Specific Recommendations:

- Downgrade one or two levels for differences in the direction of effect. See [Example Box 5.1.3.2](#).
 - If confidence intervals (CIs) are overlapping, magnitude of inconsistency is small.
 - If CIs are not overlapping, then inconsistency is larger.
 - If estimates are in the same direction, non-overlapping CIs are less critical.
- Unexplained inconsistency is undesirable: consider subgroups by population differences (by sex, race, species etc.) to explain the inconsistency.
- Reducing inconsistency may require modifying the protocol, obtaining more technical feedback, or reconsidering grouping or splitting of outcomes.

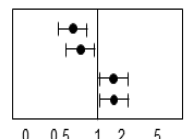
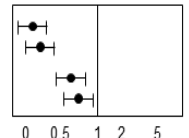
Example Box 5.1.3.2:

CONSIDER DIRECTION OF EFFECT ESTIMATES AND THE CONFIDENCE INTERVALS



Examples:

- Same direction, overlapping CIs: confidence unchanged
- Same direction, non-overlapping CIs: consider downgrading 1 level
- Different direction, non-overlapping CIs:



⁶ Studies are not rated up for consistency, because a consistent bias can be present and lead to consistent, but spurious, findings (Guyatt *et al.* 2011d)

8.1.3.3 Indirectness

Indirectness can lower confidence in the body of evidence when the population, exposure, or outcomes measured differ from those that are of most interest. The aspects of environmental health studies relevant to rating indirectness will depend on the objectives of the evaluation. Concerns about directness could apply to the relationship between a measured outcome and a health effect (i.e., upstream biomarker versus health effect), the route of exposure and the typical human exposure, or the study population and the population of interest (Guyatt *et al.* 2011c, Lohr 2012).

Specific Recommendations:

- *Downgrade one or two levels depending on the extent of the indirectness for*
 - *Upstream biomarkers instead of primary outcome measures that are more strongly linked to a health effect*
 - *Applicability of the population studied, e.g., patient population, occupational cohort. See [Example Box 5.1.3.3](#).*
- Example Box 5.1.3.3:**

OCCUPATIONAL DATA AND THE HEALTHY WORKER EFFECT

 - *For negative data (reporting a lack of a health effect with exposure to the substance) this would include the healthy worker effect*
 - *If making conclusions from occupational data for the general population consider downgrading*
 - *If making conclusions for the occupational setting from occupational data, do not downgrade*
 - *For positive data (reporting a health effect with exposure to the substance) the healthy worker effect is not relevant, as it would suggest an underestimation of the effect on the general population. Do not downgrade.*
- **For animal studies**
 - *Animal models are considered a good model for human hazard identification unless there is compelling evidence to the contrary.*
 - *Route of exposure may be an issue for experimental animal studies when the dosing regimen utilized in the study bypasses an important metabolic step with relevance to the toxicity or health effect observed. For example, intraperitoneal exposure studies circumvent digestion and metabolism associated with oral exposure, and therefore, on a case-specific basis could be considered to suffer from indirectness.*
 - *Consideration of issues of indirectness may also require modifying the protocol, obtaining more technical feedback, or reconsidering grouping or splitting of outcomes.*

8.1.3.4 Imprecision

Imprecision is the lack of certainty for an estimate of effect for a specific outcome. A precise estimate should enable the evaluator to determine whether or not there is an effect (i.e., it is different from the comparison group). Confidence intervals (CIs) of the estimates of effect provide the primary evidence used in considering the imprecision of the body of evidence (Guyatt *et al.* 2011b). Conceptually, CIs represent the range of effect estimates where the true effect plausibly lies. Sample size and power contribute to the CI estimates, and a body of evidence with many small studies decreases confidence that the results are robust.

Specific Recommendations:

- *If the width of the CIs are not sufficiently narrow, rate down by one or two levels.*
- *Consider power and statistical significance of the studies.*
- *Small sample sizes or small numbers of exposed cases can contribute to a sense of fragility to the results (large statistically significant results based on only a few cases).*
- *Substantial variability alone does not necessarily render an estimate imprecise and does not require downgrading if it is still clear that the estimate is different from that of the comparison group.*

8.1.3.5 Publication bias

Publication bias specifically pertains to the body of evidence, as selective reporting within a study is covered in risk of bias criteria addressing these limitations (Guyatt *et al.* 2011e). The structure of the systematic review literature search and screening of studies can reduce publication bias if it is comprehensive, without a language restriction, and eliminates duplicate publications to avoid double counting results. There is empirical evidence that studies with negative results (no association) are less likely to be in the published literature. Negative studies may also be affected by a “lag bias” or longer time to publication.

Specific Recommendations:

- *Author affiliations and funding source can contribute to publication bias when results are not consistent with expectations or value to the research objectives.*
- *Smaller studies or areas with few studies are more likely to be biased.*
- *A funnel plot of the magnitude of effect sizes by the precision of the estimate can visually indicate that there might be a publication bias; however, these plots are prone to error and are only part of the assessment of publication bias.*
- *Down grade only for “strongly suspected” evidence of publication bias with a maximum downgrade of one level; publication bias is*

difficult to assess an likely frequent so caution should be used in downgrading for this element.

8.1.4 Upgrade confidence rating

Upgrading confidence should be done when there is sufficient confidence in the body of evidence such that there were few concerns about the estimates of effect. Justification for upgrading the confidence in a body of evidence can be made when the evidence is stronger than could be explained by unmeasured confounding or chance findings (Guyatt *et al.* 2011f, Lohr 2012).

8.1.4.1 Large magnitude of effect

A large magnitude of effect is defined when an observed effect that is large enough that it unlikely to have occurred as a result of bias from potential confounding factors. If the estimate of the effect of an exposure on an outcome is sufficiently large, there is increased confidence in the result. There is evidence that confounding alone is unlikely to fully explain a two-fold change, and very unlikely to explain a five-fold change in effect. However, a conservative approach should be taken when rating up for a large effect size, and only done if there are no major concerns about other issues, particularly risk of bias, precision, and publication bias. Confidence intervals of large estimates should also exclude smaller effect estimates.

Specific Recommendations:

- *If there are no major concerns about effect estimates:*
 - *Rate u 1 level for two-fold change*
 - *Rate u 2 levels for five-fold change*
- *Odds ratios (ORs) require a higher threshold than relative risk estimates. When the baseline risk is high (over 40%) ORs will overestimate the effect.*

8.1.4.2 Dose-response

A dose-response relationship between level of exposure and risk of outcome can increase confidence in results if it reduces concern that results could be due to chance occurrence. By considering dose-response across a body of evidence, multiple observational human studies with varied exposure levels can contribute to an overall picture of the dose-response. Studies with multiple exposure levels, particularly experimental studies in animals and humans, are expected to display a dose-response relationship. It is important to recognize that the dose response relationship may not be monotonic and that biological plausibility should be considered in evaluating the dose-response relationship.

Specific Recommendations:

- *In observational studies, upgrade 1 level for evidence of a dose-response relationship.*
- *In experimental studies, higher standard for evidence should be used when considering whether to upgrade 1 level for evidence of a dose-response relationship.*
- *Dose response can be observed across multiple studies, with different exposure levels among the studies.*
- *Monotonic or non-monotonic dose-responses that are considered biologically plausible can provide the basis for upgrading 1 level.*

8.1.4.3 ***All plausible confounding***

When a body of evidence is potentially biased by residual confounding, which is expected to go in a direction counter to the observed effect, confidence in the results increases.

If studies find evidence of an effect:

Specific Recommendations:

- *If the residual confounding/bias is towards the null, such that hypothetical adjustment for it would actually make the estimate of effect stronger, then the confidence in the results is higher and can be upgraded.*
- Se [Example Box 5.1.4.3A](#).

Example Box 5.1.4.3A: All plausible confounding

IF STUDIES FIND EVIDENCE OF AN EFFECT AND RESIDUAL CONFOUNDING IS TOWARD THE NULL

Example from Guyatt et al. (2011f)

- Condom use prevents HIV infection, but number of sexual partners was not considered in the analysis.
- Condom users had more partners, so the effect would be stronger if it had been taken into consideration.
- Upgrade one level

If studies find no evidence of an effect:

Specific Recommendations:

- *If residual confounding would bias results away from the null (towards finding an effect), yet no effect was seen, this would increase confidence in the finding of no effect.*
- Se [Example Box 5.1.4.3B](#).

**IF STUDIES FIND NO EVIDENCE OF AN EFFECT AND RESIDUAL
CONFOUNDING IS AWAY FROM THE NULL**

Example from Guyatt et al. (2011f)

- Observational studies find no association between vaccines and autism.
- A reporting bias would be expected among parents of autistic patients being more likely to report vaccine exposure due to widely publicized, discredited studies of an association.
- So when results show no association despite this expected bias, confidence in the lack of association increases
- Upgrade one level

8.1.4.4 *Other reasons for upgrading*

There will be rare instances when the particular design features of extremely rigorous, well-conducted studies may justify rating up of confidence for other reasons. The framework allows for documenting those reasons under this category. See [Example Box 5.1.4.4](#).

Example Box 5.1.4.4: Other reasons for upgrading

Example from Guyatt et al. (2011f)

- Sigmoidoscopy is associated with reduced colon cancer mortality, but only for lesions in the range of the sigmoidoscope.
- Bias from unmeasured confounding would be the same for those within and outside the range, “considerably raising confidence in the causal effect of sigmoidoscopy.”
- Upgrade one level

8.2 If data are sufficient, consider performing meta-analysis and generating an estimate of effect within a study type (e.g., for all prospective studies)

Systematic review is often misinterpreted as synonymous with meta-analysis, where the goal is a quantitative synthesis of results from studies of the same exposure and outcome. Often these exposures and outcomes are narrowly defined from the

beginning of the systematic review to ensure that the selected studies are sufficiently similar. Typically these analyses are only performed within specific types of study design. For environmental health topics that draw on a diverse body of literature pertaining to a more broadly defined objective, quantitative synthesis by a meta-analysis will often not be appropriate. NTP anticipates that the majority of reviews will not result in a meta-analysis, but when meta-analysis is appropriate, technical experts will be involved and the methods employed will be clearly detailed within the protocol.

8.3 Combine confidence conclusions for all study types

Confidence ratings are initially set based on available study designs for a given outcome (e.g., for prospective studies separately from cross-sectional studies). For environmental health issues where protecting public health is the objective, conclusions should be based on the evidence with the highest confidence. The project-specific definition of an outcome and the grouping of biologically related outcomes used here should follow the definitions developed *a priori* in the protocol; deviations should be taken with care, justified, and documented. Project-specific explanation of the approach used for combining confidence ratings across study types should be documented.

Specific Recommendations:

- *In general, if the confidence rating for a specific outcome differs by study type, then the study type with the highest rating forms the basis of the confidence conclusion and the body of evidence (including the different study types) for the outcome and the confidence rating is used in [step 9.0](#).*
- *Otherwise, different study designs must be considered separately, and confidence conclusions should be based on the study type with the highest confidence. This study type forms the basis of the confidence conclusion and the body of evidence (including the different study types) for the outcome is used in [step 9.0](#).*
- *Confidence conclusions are not increased by considering multiple study types. In other words low confidence in the prospective studies and low confidence for the cross-sectional studies results in low confidence for the combined body of evidence for that outcome. The confidence is not increased to moderate by having multiple study types.*

8.4 Develop confidence conclusions for multiple outcomes

After confidence conclusions have been developed for a given outcome, conclusions for multiple outcomes and the entire evaluation should be developed. As stated above for multiple study types, for environmental health issues where protecting public health is the objective, conclusions should be based on the evidence with the highest confidence. Project-specific explanation of the approach used for combining confidence ratings across study types should be documented.

Specific Recommendations:

- *If the confidence rating is the same for different outcomes, then the outcome with the highest rating forms the basis of the confidence conclusion used in [step 9.0](#).*

- For biologically related outcomes (for example blood pressure and cardiovascular mortality) the reviewers should consider the outcomes in two steps:
 - Each outcome should first be considered separately, and the confidence should be stated for each health outcome separately.
 - Decisions on the relatedness or grouping of outcomes should have been documented in **step 4.2.2 Determine grouping and hierarchy of outcomes** and should be referenced. Deviations should be taken with care, justified, and documented.
 - If, in the judgment of the reviewers, the outcomes are sufficiently biologically related that they inform confidence on the overall health outcome, they should be considered together.
 - The biologically related outcomes should then be considered together and each step in **8.1.3 Downgrade confidence rating** and **8.1.4 Upgrade confidence rating** should be reconsidered for the dataset as a whole.
 - The confidence ratings of the overall body of evidence may differ from that suggested for separate outcomes. See **Example Box 5.4**.
 - It is theoretically possible that related outcomes suggest effects in opposite directions (i.e., substance Y is associated with an increase in IQ but a decrease in academic performance as determined by end of grade test scores). In this case, the available studies for each outcome may have been “consistent” when considered alone, but would be “inconsistent” when considered together and would result in a lower confidence rating of the overall body of evidence. Expert judgment may be needed to determine if related outcomes contradict each other, or if there is a plausible explanation for differences.

Example Box 5.4: confidence ratings across biologically-related

CONFIDENCE RATINGS OF THE OVERALL BODY OF EVIDENCE MAY DIFFER FROM THAT OF SEPARATE OUTCOMES

- Imprecision due to small sample size in available studies of peripheral arterial disease with compound X may lead to the conclusion of low confidence in the body of evidence for peripheral arterial disease.
- Examination of clinical cardiovascular disease (which includes peripheral arterial disease as well as coronary arterial disease, etc.) associated with compound X may lead to the conclusion of moderate confidence in the body of evidence because the available data for this more inclusive category does not suffer from imprecision.

- For unrelated outcomes, different outcomes should be considered separately, and the confidence should be stated for each health outcome separately. Confidence conclusions for “any” health effect should be based on the outcome with the highest confidence. This study type forms the basis of the confidence conclusion and the overall body of evidence. Conclusions in subsequent steps should reference whether they are specific to an individual outcome or overall for any health effect and this discussion will reflect the objectives of the evaluation.
- More complex scenarios are difficult to anticipate and should be addressed on a case-by-case basis using the principle of basing the confidence conclusions on the study type with the highest confidence.
- As described above, confidence conclusions could be increased by considering multiple related endpoints. It is unlikely that consideration of unrelated endpoints would affect the confidence in the overall conclusions. For example, low confidence that substance

is a dermal irritant does not affect the confidence that it is associated with decreased renal function.

Table 4: Evidence of Health Effects Descriptors

Table 5: Relationship between Bradford Hill Criteria (Hill 1965) of causality in the NTP approach for upgrading and downgrading confidence in a body of evidence (based on the GRADE approach as described in Schünemann *et al.* 2011)

9.0 TRANSLATE CONFIDENCE RATINGS INTO EVIDENCE OF HEALTH EFFECT

The confidence in the evidence is distinct from the study findings. The evidence of health effects conclusions reflects both the overall confidence in the association between exposure to the substance and the outcome (effect or no effect) and the direction of the effect (toxicity or no toxicity).

9.1 Evidence of Health Effects Descriptors

The NTP approach uses 4 terms to describe the evidence of health effects; these descriptors reflect both the confidence in the body of evidence for a given outcome and the direction of effect. There are 3 descriptors (“sufficient”, “limited”, and “inadequate”) to indicate confidence that exposure to the substance is associated with a health effect; and a fourth designation (evidence of no health effect(s)) to indicate confidence that the substance is not associated with a health effect (see [Table 4](#) for detailed definitions). Although the conclusions describe associations, a causal relationship is implied and the ratings describe evidence of health effects in terms of confidence in the association or the estimate of effect determined from the body of evidence (see [Table 5](#) for discussion of the NTP approach and the Hill criteria for causation).

Descriptors	Definition
Sufficient	There is high confidence in the body of evidence for an association between exposure to the substance and the health outcome(s).
Limited	There is moderate confidence in the body of evidence for an association between exposure to the substance and the health outcome(s).
Inadequate	There is low or very low confidence in the body of evidence for an association between exposure to the substance and the health outcome(s) or no data are available.
Evidence of no health effect	There is high confidence in the body of evidence that exposure to the substance is not associated with the health outcome(s).

Hill Criteria	Consideration in the NTP approach
Strength	Considered in upgrading the confidence in the body of evidence for large magnitude of effect and downgrading confidence for imprecision
Consistency	Considered in downgrading confidence in the body of evidence for inconsistency
Temporality	Considered in initial confidence ratings by study design, for example experimental studies are rated “High” because of the increased confidence that exposure preceded outcome
Biological gradient	Considered in upgrading the confidence in the body of evidence for evidence of a dose-response relationship
Biological plausibility	Considered in downgrading the confidence in the body of evidence for indirectness ; also in examining non monotonic dose-response relationships
Experimental evidence	Considered in downgrading for risk of bias and initial confidence ratings by study design

Note that more recent interpretations of the Bradford Hill Criteria of causality generally do not consider 3 of the original 9 criteria (biological plausibility covers coherence and specificity, and analogy is considered of limited utility (Szklo and Nieto 2007).

9.2 Consider evidence of effect or lack of effect

The determination of overall confidence for a given outcome is translated into evidence of health effects by considering the direction of the effect (i.e., is there evidence of a health effect or not; see [Table 6A](#) and [Table 6B](#)). As discussed in [Section 8.0 Rate the Confidence in the Body of Evidence](#), a single, well conducted study may provide evidence of toxicity or a health effect associated with exposure to the substance in question. In contrast, evidence that exposure to the substance is not associated with health outcomes would require a larger body of evidence. First, no single study can test all potential outcomes and, therefore, a larger number of studies are required to support an initial finding of a lack of an effect. Second, this health protective approach when scientific judgments are made is consistent with NTP’s objective to identify potential environmental hazards to safeguard public health. The determination of evidence of effect or lack of effect is made separately within the human and experimental animal data sets.

9.2.1 *Evidence of a health effect or health effects*

When there is evidence of a health effect, high confidence in a body of evidence directly translates into the conclusion of sufficient evidence of an association between exposure to the substance and the health effect. Similarly, moderate confidence translates into the conclusion of limited evidence of an association between exposure to the substance and the health effect. The two lower levels of confidence (low and very low) are reflected in the conclusion of inadequate evidence of an association between exposure to the substance and the health effect.

Table 6A: NTP Procedure for Determining Evidence of Health Effects Conclusions
Table 6B: NTP Procedure for Determining Evidence of a Lack of a Health Effects

Confidence in the Body of Evidence	Direction (effect or no effect)	Evidence of Health Effect
(++++) high	Health effect	Sufficient
(+++) moderate	Health effect	Limited
(++) low	Health effect	Inadequate
(+) very low	Health effect	Inadequate

9.2.2 Evidence of no effect or lack of toxicity

Because of the inherent difficulty in proving a negative, the conclusion of evidence of no toxicity is only possible when there is high confidence in the body of evidence. When there is no evidence of a health effect, high confidence in a body of evidence translates into the conclusion of evidence of no health effect or the lack of an association between exposure to the substance and the health effect. When there is no evidence of a health effect, all three lower levels of confidence (moderate, low, and very low) result in the conclusion of inadequate evidence.

Confidence in the Body of Evidence	Direction (effect or no effect)	Evidence of Health Effect
(++++) high	No effect	Evidence of no health effect
(+++) moderate	No effect	Inadequate
(++) low	No effect	Inadequate
(+) very low	No effect	Inadequate

9.2.3 Health effects conclusions for multiple outcomes

As noted in step **8.4 Develop confidence conclusions for multiple outcomes**, confidence conclusions should be developed for individual outcomes or groups of biologically related outcomes. Similarly, these confidence conclusions should be translated into health effects conclusions for the individual health effects and the combined (biologically related) effects.

10.0 INTEGRATE EVIDENCE TO DEVELOP HAZARD IDENTIFICATION CONCLUSIONS

Table 7A: NTP Procedure for Reaching Hazard ID Conclusions through Considering Human and Experimental Animal Evidence

To reach the hazard identification conclusion, the evidence for a health effect from each of the evidence streams is combined in the final step of the evidence assessment process. Hazard identification conclusions can be reached on individual outcomes (health effects) or groups of biologically related outcomes, as appropriate, based on the evaluation's objectives and the available data. The rationale for such conclusions are documented as the evidence is combined within and across evidence streams and the conclusions should be clearly stated as to which outcomes are incorporated in each conclusion. The four hazard identification conclusion categories are described below with their equivalent Globally Harmonized System of Classification and Labeling of Chemicals category.

- Known to be a hazard to humans [equivalent to Category 1A GHS]
- Presumed to be a hazard to humans [equivalent to Category 1B GHS]
- Suspected to be a hazard to humans, [equivalent to Category 2 GHS]
- Not classifiable or not identified to be a hazard to humans

10.1 Conclusions based on combination of evidence streams

The evidence streams (human studies and non-human animal studies) have remained separate through all previous steps in the NTP Approach. In step 7 they are integrated along with other relevant data⁷ such as supporting evidence from mechanistic studies. First, the evidence of health effects conclusions for human data from [step 9.0](#) ("sufficient", "limited", or "inadequate") are considered together with the evidence of health effects conclusions for non-human data to reach one of four hazard identification conclusions as outlined in [Table 7A](#).

If the human evidence conclusion is sufficient the hazard identification conclusion is "known" based on the human data alone. If the human evidence conclusion is limited, the hazard identification conclusion depends on the strength of the non-human evidence. The hazard identification conclusion is "presumed" if the non-human evidence conclusion is sufficient or "suspected" if the non-human evidence conclusion is limited or inadequate. If the human evidence conclusion is inadequate, the hazard identification conclusion again depends on the strength of the non-human evidence. The hazard identification conclusion is "suspected" if the non-human evidence conclusion is sufficient or "not classifiable" if the non-human evidence conclusion is limited or inadequate.

⁷ See http://oehha.ca.gov/multimedia/green/pdf/GC_Regtext011912.pdf for definition and discussion of "Other relevant data"; in brief it refers to non-endpoint data, including chemical, physical, biochemical, biological or other data that may indicate a chemical substance may contribute to adverse effects.

Table 7B: NTP Procedure for Reaching Hazard ID Conclusions with Consideration of Other Relevant Data (e.g., supporting evidence from mechanistic Studies)

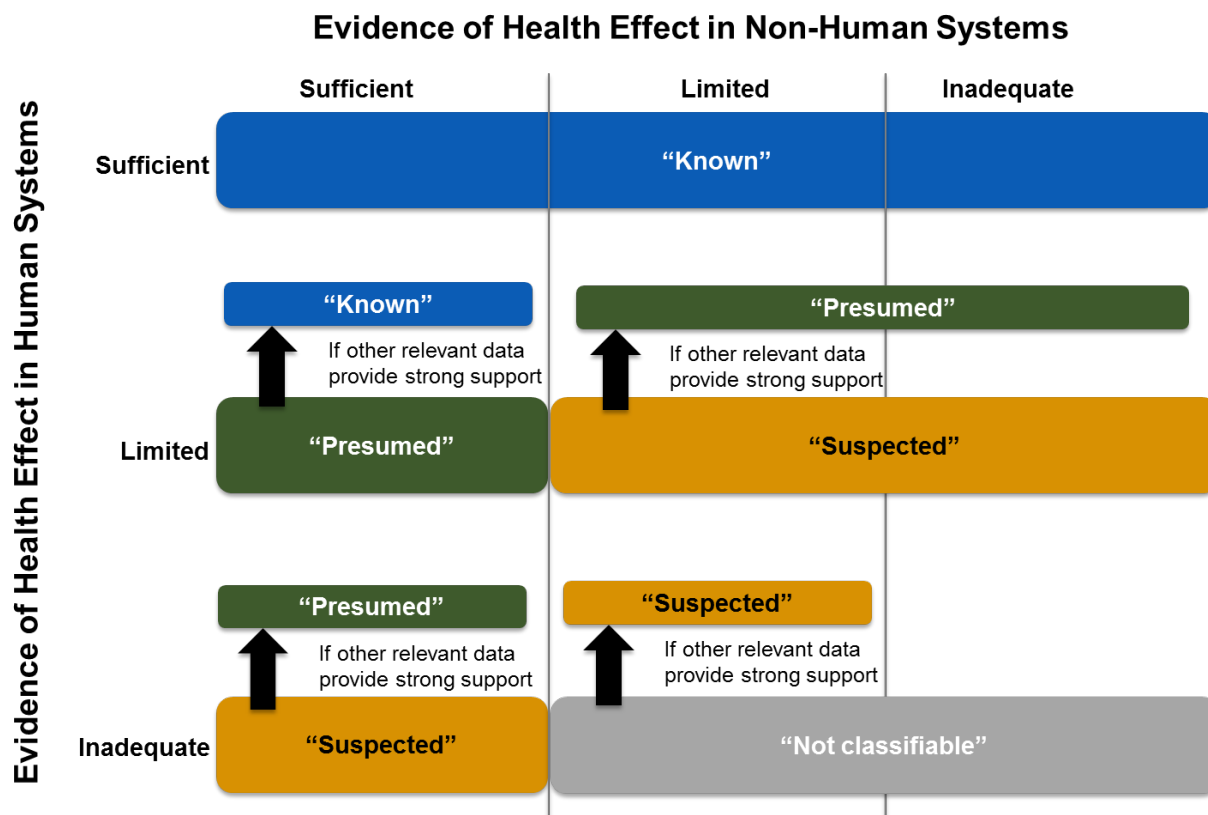
		Evidence of Health Effect in Non-Human Systems		
		Sufficient	Limited	Inadequate
Evidence of Health Effect in Human Systems	Sufficient	"Known"		
	Limited	"Presumed"	"Suspected"	
	Inadequate	"Suspected"	"Not classifiable"	

■ "Known" Category 1A GHS
■ "Presumed" Category 1B GHS
■ "Suspected" Category 2 GHS
■ "Not classifiable" or not identified to be hazard to humans

10.2 Consider mechanistic, *in vitro*, or other supporting data

The hazard identification conclusion that could be derived by integrating the human and non-human animal streams should also consider other relevant evidence such as mechanistic data, *in vitro* data, and evidence on upstream indicators of a health effect that might otherwise be overlooked.

As outlined in [Table 7B](#), strong supporting evidence may raise the level of the hazard identification conclusion initially derived in [step 10.1](#). Note that mechanistic or supporting evidence is not required to reach a hazard identification conclusion of "known" if human evidence is sufficient. If the hazard identification conclusion was "presumed" based on the human and non-human data, strong support from other relevant data may result in an upgrade conclusion of "known." If the hazard identification conclusion was "suspected" based on the human and non-human data, strong support from other relevant data may result in an upgraded conclusion of "presumed." If the human evidence conclusion is inadequate and the non-human evidence is limited, consideration of other relevant data can be used to reach a hazard identification conclusion of "suspected."



ACKNOWLEDGMENTS

The NTP's approach for evidence assessment was developed by the Office of Health Assessment and Translation (OHAT) and Office of Liaison, Policy and Review (OLPR), within the Division of the National Toxicology Program at the NIEHS. Strong support for undertaking this project was provided by the NTP Board of Scientific Counselors, NTP Executive Committee, public, and other stakeholders (see <http://ntp.niehs.nih.gov/go/9741> for meeting minutes). In developing this methodology, the NTP considered authoritative sources on systematic review including, but not limited to, the Agency for Healthcare Research and Quality (AHRQ) (AHRQ 2012), The Cochrane Collaboration (Higgins and Green 2011), the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (Guyatt *et al.* 2011a), and the Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Studies (CAMARADES, see <http://www.camarades.info/>). Additional technical expertise was provided on portions of this method by experts affiliated with these groups including Lisa Bero, Director, San Francisco Branch, United States Cochrane Center at UC San Francisco; Gordon Guyatt, Co-chair, GRADE working group, McMaster University; Malcolm Macleod, CAMARADES Centre, University of Edinburgh; Karen Robinson, Co-Director, AHRQ Evidence-Based Practice Center, The Johns Hopkins Bloomberg School of Public Health; Holger Schünemann, Co-chair, GRADE working group, McMaster University; and Tracey Woodruff, Director, Program on Reproductive Health and the Environment, UC San Francisco.

REFERENCES

- AHRQ. 2012. Grading the Strength of a Body of Evidence When Assessing Health Care Interventions: An Update (Draft Report). Available at <http://effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=1163> [accessed July 30, 2012].
- ATSDR. 2012. *The Future of Science at ATSDR: A Symposium*, Atlanta, GA, U.S. Department of Health and Human Services (DHHS) Agency for Toxic Substances and Disease Registry (ATSDR).
- Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. 2011. GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology* 64(4): 401-406.
- EFSA. 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. Available at: <http://www.efsa.europa.eu/en/efsajournal/pub/1637.htm> [accessed January 18, 2012]. *EFSA Journal* 8(6): 1637 [1690 pp.].
- Foster PM. 2009. *Explanation of Levels of Evidence for Reproductive System Toxicity*. National Toxicology Program. <http://ntp.niehs.nih.gov/go/18711>. Research Triangle Park, NC: US Department of Health and Human Services.
- Germolec D. 2009. *Explanation of Levels of Evidence for Immune System Toxicity*. National Toxicology Program. <http://ntp.niehs.nih.gov/go/9399> Research Triangle Park, NC: US Department of Health and Human Services.
- Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, Debeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schunemann HJ. 2011a. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology* 64(4): 383-394.
- Guyatt G. 2012. Tools to Assess Risk of Bias in Cohort Studies, McMaster University: Available at: <http://www.evidencepartners.com/resources/> [accessed July 13, 2012].
- Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, Atkins D, Kunz R, Montori V, Jaeschke R, Rind D, Dahm P, Akl EA, Meerpohl J, Vist G, Berliner E, Norris S, Falck-Ytter Y, Schunemann HJ. 2012. GRADE guidelines 11-making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. *Journal of clinical epidemiology*.
- Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Devereaux PJ, Montori VM, Freyschuss B, Vist G, Jaeschke R, Williams JW, Jr., Murad MH, Sinclair D, Falck-Ytter Y, Meerpohl J, Whittington C, Thorlund K, Andrews J, Schunemann HJ. 2011b. GRADE guidelines 6. Rating the quality of evidence--imprecision. *Journal of clinical epidemiology* 64(12): 1283-1293.
- Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Falck-Ytter Y, Jaeschke R, Vist G, Akl EA, Post PN, Norris S, Meerpohl J, Shukla VK, Nasser M, Schunemann HJ. 2011c. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *Journal of clinical epidemiology* 64(12): 1303-1310.
- Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Glasziou P, Jaeschke R, Akl EA, Norris S, Vist G, Dahm P, Shukla VK, Higgins J, Falck-Ytter Y, Schunemann HJ. 2011d. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *Journal of clinical epidemiology* 64(12): 1294-1302.
- Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y, Williams JW, Jr., Meerpohl J, Norris SL, Akl EA, Schunemann HJ. 2011e. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *Journal of clinical epidemiology* 64(12): 1277-1282.
- Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, Atkins D, Kunz R, Brozek J, Montori V, Jaeschke R, Rind D, Dahm P, Meerpohl J, Vist G, Berliner E, Norris S, Falck-Ytter Y, Murad MH, Schunemann HJ. 2011f. GRADE guidelines: 9. Rating up the quality of evidence. *Journal of clinical epidemiology* 64(12): 1311-1316.
- Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams JW, Jr., Atkins D, Meerpohl J, Schunemann HJ. 2011g. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *Journal of Clinical Epidemiology* 64(4): 407-415.
- Higgins J, Green S, eds. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 [updated March 2011]: The Cochrane Collaboration. Available at: www.cochrane-handbook.org [accessed January 18, 2012].
- Hill AB. 1965. The Environment and Disease: Association or Causation? *Proc R Soc Med* 58: 295-300.
- Khan K, ter Riet G, Glanville J, Sowden A, Kleijnen J, eds. 2001. *Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guidance for those Carrying Out or Commissioning Reviews (CRD Report Number 4) (2nd edition)*. York (UK): NHS Centre for Reviews and Dissemination: University of York.

- Lohr KN. 2012. Grading the Strength of Evidence. The Agency for Healthcare Research and Quality (AHRQ) Training Modules for Systematic Reviews Methods Guide, Available at: <http://www.effectivehealthcare.ahrq.gov/index.cfm/slides/?pageAction=displaySlides&tk=18> [accessed July 13, 2012].
- National Research Council. 2011. Committee to Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Available at: http://www.nap.edu/openbook.php?record_id=13142 [accessed July 30, 2012]. *The National Academies Press*.
- Schünemann H, Hill S, Guyatt G, Akl EA, Ahmed F. 2011. The GRADE approach and Bradford Hill's criteria for causation. *J Epidemiol Community Health* 65(5): 392-395.
- Schünemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, P. G, Guyatt GH, on behalf of the Cochrane Applicability and Recommendations Methods Group. 2012. Chapter 12: Interpreting results and drawing conclusions. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]*. Higgins JPT, Green S, eds., The Cochrane Collaboration, 2011. Available at www.cochrane-handbook.org. [accessed July 13, 2012].
- Szklo M, Nieto FJ. 2007. *Epidemiology: Beyond the basics* 2nd, Sudbury, MA: Jones and Bartlett Publishers.
- Viswanathan M, Ansari MT, Berkman ND, Chang S, Hartling L, McPheeters M, Santaguida PL, Shamliyan T, Singh K, Tsertsvadze A, Treadwell JR. 2012. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. In *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Agency for Healthcare Research and Quality (AHRQ) Methods for Effective Health Care, AHRQ Publication No. 12-EHC047-EF. Rockville (MD). Available at: www.effectivehealthcare.ahrq.gov/ [accessed July 13, 2012].
- Woodruff TJ, Sutton P. 2011. An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. *Health affairs (Project Hope)* 30(5): 931-937.

APPENDIX C

The NTP's Approach for Reaching Conclusions for Literature-Based Evidence Assessments

NTP Board of Scientific Counselors Working Group Meeting
Raleigh Marriott Crabtree Valley, 4500 Marriott Drive, Raleigh, NC

August 28-29, 2012

CHARGE

to obtain feedback on the NTP's proposed approach for reaching conclusions for literature-based evidence assessments

BACKGROUND

The Draft NTP Approach for Reaching Conclusions for Literature-Based Evidence Assessments presents a methodology to use the information gathered by transparent systematic review process to reach hazard identification conclusions.

FORMAT FOR MEETING

- 1) Presentation on the topic to be discussed including relevant background information and examples when applicable

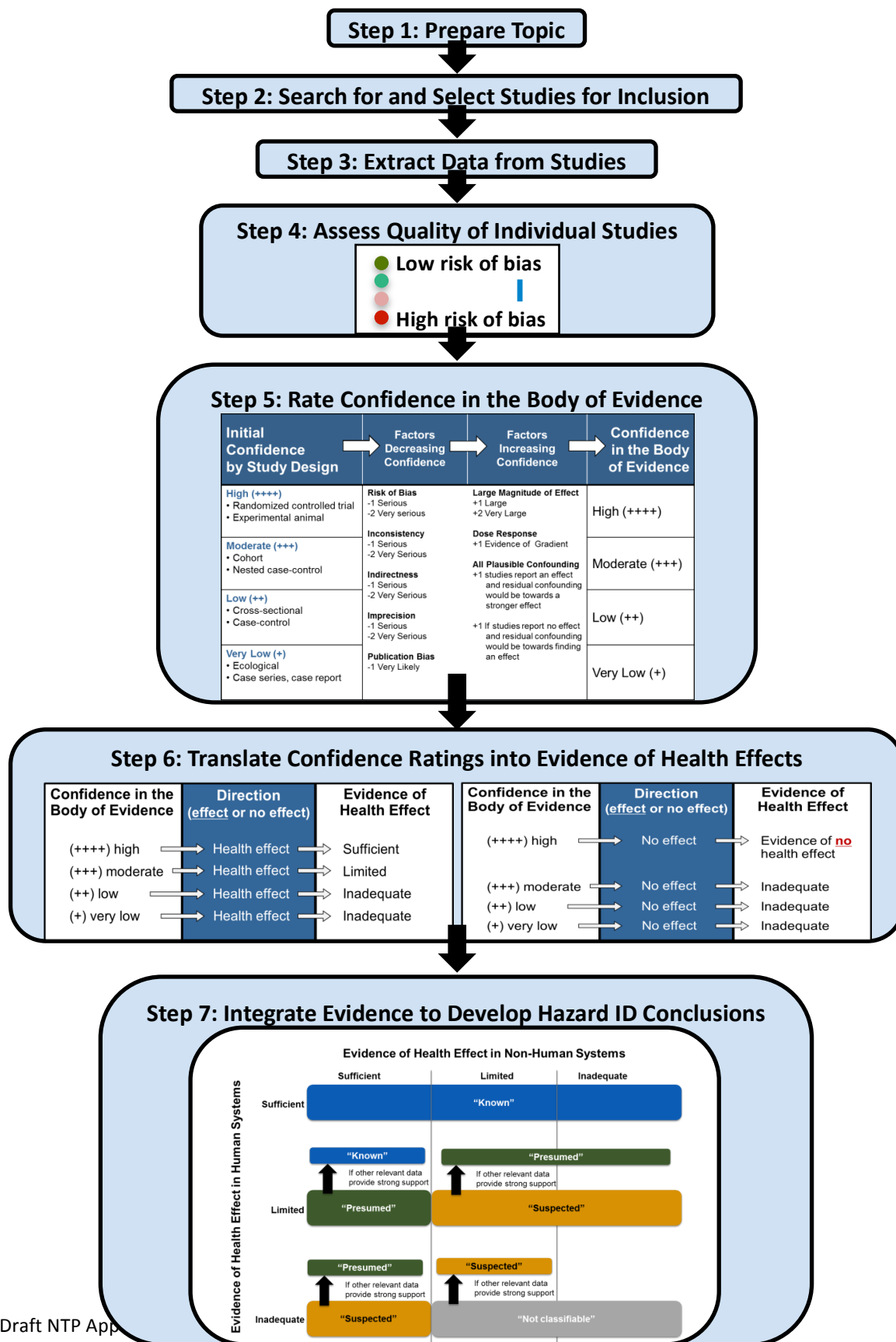
Order of presentations:

- 1st talk:** Systematic review background and initial steps of the NTP approach
 - 2nd talk:** Step of the approach – Assess the quality or risk of bias of individual studies
(specific area for consultation)
 - 3rd talk:** Step of the approach – Rate the confidence in the body of evidence *(specific area for consultation)*
 - 4th talk:** Step of the approach – Translate confidence ratings into evidence of health effects
(specific area for consultation)
 - 5th talk:** Step of the approach – Integrate evidence to develop hazard identification conclusions *(specific area for consultation)*
- 2) Working group discussion and consideration of specific areas for comment
 - Comments on NTP's proposed approach
 - Recommended modifications, as needed

The following specific areas for comment relate to steps in the NTP Approach for Reaching Conclusions for Literature-Based Evidence Assessments for which we are consulting the working group. NTP has highlighted major points within each step to focus the working group's discussion that are arranged below under headings that match the steps in the process (see Figure 1 of the Draft NTP Approach for Reaching Conclusions for Literature-Based Evidence Assessments for a description of all 7 steps). NTP is not asking for comment on steps 1-3 because the first three steps

describe standard systematic review procedures consistent with AHRQ and other methodologies. The specific areas for consultation start with step 4.

Figure 2. NTP Framework for Conducting Literature-Based Evidence Assessment



STEP 4: ASSESS THE QUALITY OR RISK OF BIAS OF INDIVIDUAL STUDIES

- 1) **NTP is proposing to use a modified list of 18 risk of bias questions to assess the quality of individual studies** (see Table 1 of the Draft NTP Approach document - also copied below - for a list of all 18 questions and application of individual questions to specific study-types). The questions would be answered for each outcome in a study.
 - a. Do these questions cover the aspects of study design and execution that are relevant for assessing risk of bias?
 - b. If not, what's missing or unclear?
- 2) **Default rating for an outcome where information is not reported in a paper is "Probably High" risk of bias**
- 3) **Default rating for "emerging" endpoints/exposure assessment techniques is "Probably Low" risk of bias**
- 4) **Of the 18 questions, 4 are designated as "major" risk of bias criteria** (shaded in Table 1 below)
- 5) **If an outcome-specific result is rated "Definitely High" risk of bias for 2 or more major elements, the study would be excluded for that outcome.**
 - a. Should these excluded studies be considered in a stratified or sensitivity analysis? If so, when in the process should this be done?
 - b. Should these studies be reconsidered as supporting material, or do the quality concerns prevent further consideration?

Table 1: NTP Risk of Bias Questions

Study Design-specific Risk of Bias Questions		Experimental Animal	RCT	Cohort	Case control	Cross sectional	Case Series
Types of Bias	Questions						
Selection	Was treatment or exposure adequately randomized?	x	x				
	Was treatment or exposure allocation adequately concealed?	x	x				
	Was the subject recruitment strategy uniform across study groups?		x	x			
	Is the comparison group appropriate, including similar baseline characteristics and having the exposed and non-exposed subjects drawn from the same population?	x	x	x	x	x	x
	Does the study design adjust/control for important confounding and modifying variables?	x	x	x	x	x	x
Performance	Did researchers adjust/control for other exposures or interventions that may bias results?	x	x	x	x	x	x
Attrition	In RCT, animal, or cohort studies: does the length of follow-up differ between groups?						
	In case-control studies: is the time period between exposure/intervention and outcome the same for cases and controls?	x	x	x	x		
	Was the attrition rate uniformly low?	x	x	x	x		
	Is the analysis conducted on an intention-to-treat basis?		x	x			
Detection	Was follow-up long enough to assess the outcome of interest?	x	x	x			
	Can we be confident that the outcome did not precede exposure?	x	x	x			x
	Were outcome assessors blinded to the exposure or intervention status of participants?	x	x	x	x	x	x
	Is inclusion/exclusion criteria measured reliably, implemented consistently?	x	x	x	x	x	x
	Are confounding variables assessed using reliable measures, implemented consistently?			x	x	x	x
	Are data analyses appropriate, performed with reliable tests, implemented consistently?	x	x	x	x	x	x
	Can we be confident in the exposure assessment?	x	x	x	x	x	x
Reporting	Can we be confident in the outcome assessment?	x	x	x	x	x	x
	Are outcomes pre-specified by the researchers? Are all pre-specified outcomes reported?	x	x	x	x	x	x
Other							

An "x" at the intersection of a question-row and study-type-column indicates the risk of bias question applies to that study type. A shaded box is used to identify that the question is considered a "major" question for that study type.

STEP 5 RATE THE CONFIDENCE IN THE BODY OF EVIDENCE

- 1) Initial confidence rating is based on study design
- 2) Confidence ratings are decreased for concerns about the study or its findings (i.e., risk of bias of the body of evidence, inconsistency, indirectness, imprecision, publication bias)
- 3) Confidence ratings are increased for strengths in the study or its findings (i.e., large magnitude of effect, dose-response, all plausible confounding)
 - a. In observational studies, evidence of a dose response is considered a reason to upgrade the confidence in the body of evidence 1 level. Is this approach reasonable?
 - b. Because experimental studies with multiple exposure levels are expected to display a dose-response relationship, a higher standard of evidence is suggested when considering whether to upgrade 1 level for evidence of a dose-response relationship. Is this approach reasonable?
- 4) Conclusions across study designs or multiple outcomes are based on the evidence with the highest confidence
 - a. Can confidence be increased by consistent findings from multiple study types?
 - b. Can confidence be increased by consistent findings across multiple species?
 - c. Can confidence be increased by consistent findings from multiple populations?
- 5) The confidence conclusions for biologically related outcomes may be assessed for each outcome separately and then reassessed after combining data for all the related outcomes. The overall confidence rating for combined outcomes can differ from that of the individual outcome ratings.
- 6) There are more potential factors by which confidence in body of evidence can be decreased than increased. Is this approach appropriate, or is it imbalanced considering health protective goal?

Table 3: Schematic to Develop Confidence Rating for the Body of Evidence (copied here for reference, see the Draft NTP Approach document for additional details)

Initial Confidence by Study Design	Factors Decreasing Confidence	Factors Increasing Confidence	Confidence in the Body of Evidence
High (++++) • Randomized controlled trial • Experimental animal	Risk of Bias -1 Serious -2 Very serious	Large Magnitude of Effect +1 Large +2 Very Large	High (++++)
Moderate (+++) • Cohort • Nested case-control	Inconsistency -1 Serious -2 Very Serious	Dose Response +1 Evidence of Gradient	Moderate (+++)
Low (++) • Cross-sectional • Case-control	Indirectness -1 Serious -2 Very Serious	All Plausible Confounding +1 studies report an effect and residual confounding would be towards a stronger effect	Low (++)
Very Low (+) • Ecological • Case series, case report	Imprecision -1 Serious -2 Very Serious	+1 If studies report no effect and residual confounding would be towards finding an effect	Very Low (+)
	Publication Bias -1 Very Likely		

STEP 6: TRANSLATE CONFIDENCE RATINGS INTO EVIDENCE OF HEALTH EFFECTS

- 1) Evidence of health effects conclusions reflect confidence in the body of evidence
 - **High confidence** Sufficient evidence
 - **Moderate confidence** Limited evidence
 - **Low or very low confidence** Inadequate evidence
- 2) A conclusion of evidence of no health effect requires high confidence in the body of evidence
- 3) The outcome with the highest evidence of health effect conclusion moves forward for hazard identification labeling

Tables 6A, B copied here for reference, see the Draft NTP Approach document for additional details.

Table 6A: NTP Procedure for Determining Evidence of Health Effects Conclusions

Confidence in the Body of Evidence	Direction (effect or no effect)	Evidence of Health Effect
(+++++) high	Health effect	Sufficient
(+++)+ moderate	Health effect	Limited
(++) low	Health effect	Inadequate
(+) very low	Health effect	Inadequate

Table 6B: NTP Procedure for Determining Evidence of a Lack of a Health Effects

Confidence in the Body of Evidence		Direction (<u>effect</u> or no effect)		Evidence of Health Effect
(++++) high	→	No effect	→	Evidence of no health effect
(+++) moderate	→	No effect	→	Inadequate
(++) low	→	No effect	→	Inadequate
(+) very low	→	No effect	→	Inadequate

STEP 7: INTEGRATE EVIDENCE TO DEVELOP HAZARD IDENTIFICATION CONCLUSIONS

1) **Four hazard identification conclusions**

- **Known** to be a hazard to humans
- **Presumed** to be a hazard to humans
- **Suspected** to be a hazard to humans
- **Not classifiable or not identified** to be a hazard to humans

2) **Proposed approach for consideration of other relevant data (e.g., supporting evidence from mechanistic studies) enables upgrading of hazard label**

3) **Proposed approach uses the same hazard categories as the Globally Harmonized System of Classification and Labeling of Chemicals (GHS)**

- a. Is there any downside to using the same labels for hazard ID as the Globally Harmonized System of Classification and Labeling of Chemicals (GHS)?

Table 7B: NTP Procedure for Reaching Hazard ID Conclusions with Consideration of Other Relevant Data (e.g., supporting evidence from mechanistic studies) (Table 6B copied here for reference, see the Draft NTP Approach document for additional details)

